INTERNATIONAL WORKSHOP

Wednesday 16 November, 2016

Meeting rooms Paris and Trocadéro
8, rue de la Croix-Jarry, 75013 Paris

POPULATION HEALTH INTERVENTION RESEARCH

«Process evaluation of population health intervention research:

A complement or an alternative contribution to randomized controlled trial? »

aviesan alliance nationale pour les sciences de la vie et de la santé

Instituts thématiques

Inserm Institut national de la santé et de la recherche médicale

anrs France REcherche Nord & sud Sida-hiv Hépatites Agence autonome de l'Inserm

INSTITUT NATIONAL DU CANCER

Institut de Recherche en Santé Publique IReSP

SUMMARY

CONTACTS

François Alla
Francois.alla@iresp.net

Pierre Arwidson
Pierre.ARWIDSON@santepubliquefrance.fr

Pierre Blaise
pierre.blaise@ars.sante.fr

Christopher Bonell
christopher.bonell@spi.ox.ac.uk

Emmanuel Bonnet
emmanuel.bonnet@ird.fr

Isabelle Boutron
isabelle.boutron@bch.aphp.fr

Linda Cambon
linda.cambon@ehesp.fr

Rona Campbell
rona.campbell@bristol.ac.uk

Frank Chauvin
frank.chauvin@icloire.fr

Geneviève Chêne
wanida.pellegrin@inserm.fr

Christine Chomienne
cchomienne@institutcancer.fr

François Dabis
Francois.Dabis@isped.u-bordeaux2.fr

Nancy Edwards
nedwards@uottawa.ca

Christine Ferron
christine.ferreon@fnes.fr

Marie-Renée Guével
Marie-Renee.Guevel@ehesp.fr

Nadir Kellou
nadirkellou@yahoo.fr

Joelle Kivits
joelle.kivits@univ-lorraine.fr

Anthony Lacouture
anthony.lacouture@ehesp.fr

Thierry Lang
thierry.lang@univ-tlse3.fr

Susan Michie
s.michie@ucl.ac.uk

Laetitia Minary
laetitia.minary@gmail.com

Graham Moore
mooreg@cardiff.ac.uk

Grégory Ninot
gregory.ninot@univ-montp1.fr

Kareen Nour
kareen.nour.agence16@ssss.gouv.qc.ca

Jeanine Pommier
jeanine.pommier@ehesp.fr

Louise Potvin
louise.potvin@umontreal.ca

Lehana Thabane
thabanl@mcmaster.ca

Julie Charlesworth
julie@atreeoflifesciences.com

WELCOME ADRESSES AND PRESENTATION OF PROCESS EVALUATION

Process evaluation aims "to assess fidelity and quality of implementation, clarify causal mechanisms and identify contextual factors associated with variation in outcomes." (Craig et al. 2008). The process evaluation is a key issue in the evaluation approach. It is beyond conclude whether an intervention is effective or not, but why it is effective, how it is effective and for whom it is effective. An answer to these questions is required especially for transferability and generalization processes. In this context, the recent MRC guidance (BMJ 2015) represents a key milestone. We now need to develop methods, tools and practical guidance for researchers who want to implement this guidance. We need also to clarify some underlying paradigms and to operationalize the overall research approach, from the conceptualization to the dissemination of an intervention.

The national coordinated action for intervention research (ACRISP) aims to promote the sharing of experience between researchers, practitioners and policy-makers; to encourage conceptual and methodological reflections; and to make proposals in terms of organization of research, regulation and funding. Process evaluation is one of our focuses.

After a year of operation, we wanted to go further and have organized this workshop bringing together the world's leading experts on the subject. Our objective is to promote exchanges between researchers from different disciplines. Indeed, grasping the complexity requires an interdisciplinary approach. We would also encourage the sharing of experiences between researchers from various fields, i.e. clinical research, health services research, population health intervention research. We believe we will progress by learning from each other. This workshop will not solve all issues, but we hope it can contribute. Its goal is also to bring together researchers interested in the conceptual and methodological developments for future exchanges and even partnerships. We wish you an excellent and profitable day of interactions!

PRESENTATION OF THE WORKSHOP

CONTEXT

One of the objectives of population health intervention research (PHIR) is to demonstrate the effectiveness of interventions acting on the distal and proximal determinants of health. However, due to their complex nature, the evaluations of such interventions cannot be limited to the demonstration of their effectiveness, but must also examine the mechanisms of action underlying this efficiency. It implies to explore the "black box" of interventions, their mechanisms and the interactions between context and action. This includes not only exploring the efficiency of interventions, but understanding why, how, for whom, to what extent and under what conditions interventions are working, or not. This exploratory approach is essential in order to consider the sustainability and the transfer of interventions that have demonstrated their effectiveness in population health.

In this context, the Medical Research Council Guidance (Craig et al., 20001; 20082) is stressing recommendations to guide researchers in designing, developing and evaluating complex health interventions, more specifically process evaluation (Moore, 2015)3. However questions remain unanswered regarding those methodological guidelines:

- Might process evaluation be nested in a trial or is it an alternative design?
- Which methods can be mobilized?
- What is the temporality of process evaluation (competitive, sequential…)?
- How to report the results produced, so they can participate in programs transfer?

1 Campbell, Michelle et al. "Framework for Design and Evaluation of Complex Interventions to Improve Health." BMJ : British Medical Journal 321.7262 (2000): 694–696.

2 Craig, Peter et al. "Developing and Evaluating Complex Interventions: The New Medical Research Council Guidance." The BMJ 337 (2008): a1655. PMC. Web. 4 July 2016.

3 Moore, G. F., Audrey, S., Barker, M., Bond, L., Bonell, C., Hardeman, W., Moore, L., O'Cathain, A., Tinati, T., Wight, D., and Baird, J. (2015), Process evaluation of complex interventions: Medical Research Council guidance. British Medical Journal ; 2015 : h

OBJECTIVES

By giving French researchers the opportunity to meet leading international intervention researchers, this workshop is a unique opportunity to shed light and provide answers on the issues that are currently structuring the field of intervention research. The final objectives of this seminar are to provide written recommendations on process evaluation and to publish an article in a scientific peer reviewed high ranking journal.

Three working groups will be held throughout the day. In each working group, a moderator will previously send materials (articles), prepare a brief presentation of key issues to debate, and three or four discussants will present concrete examples. These three working groups will all be followed by an open and plenary discussion. Then a moderator will conclude the roundtable.

THEMES OF THE WORKING GROUPS

1. The place of theory into process evaluation: does it highlight the role of mechanisms? Should process evaluation be theory-driven? How? How to consider what might be anticipated in the program-theory? What are the current logic models and frameworks that are combining theory and process of interventions?

2. The place of pilot studies in process evaluation: objectives, contribution (co-construction of a theory, validation/invalidation of a planned theory, evaluation of the mechanisms, pilot studies to contrast the effects of context or to test different modalities of interventions (in terms of feasibility, recruitment, inclusion, participation, etc.)?

3. Which methodologies might be combined in process evaluation (qualitative, quantitative, mixed-method, realist approach), and how?

PROGRAM

9.30 - 10.00	***Welcome coffee***
10.00 - 10.15	**Opening** **Christine Chomienne,** Director for Research and Innovation, National Cancer Institute (INCa)
10.15 - 10.30	**Introduction on the practical aspects of the workshop.** François Alla, Deputy-Director of IReSP
10.30 - 12.30	**Session 1 : The place of theory into process evaluation: does it highlight the role of mechanisms? Should process evaluation be theory-driven? How? How to consider what might be anticipated in the program-theory? What are the current logic models and frameworks that are combining theory and process of interventions?**
	Session chairs **Graham Moore, Cardiff University and Pierre Arwidson, Santé Publique France**
15	**Introduction.** Graham Moore, Cardiff University
10	**Behaviour change techniques and their mechanisms of action**, Susan Michie, University College London
10	**A Need of Integrative and Comprehensive Health Intervention Ontology for Intervention Research,** Grégory Ninot, Université de Montpellier
10	**Socioecological theory-based interventions: How to evaluate the effectiveness of their mechanisms?**, Nadir Kellou, Collège Universitaire de médecine générale, Faculté de médecin de Lyon-est, Université Claude Bernard Lyon 1
40	***Discussion***
20	***Questions of the public***
15	**Conclusion/Overview,** Pierre Arwidson, Santé Publique France
12..30 - 13.30	***Lunch break***
13.30 - 15.30	**Session 2 : The place of pilot studies in process evaluation: objectives, contribution (co-construction of a theory, validation/invalidation of a planned theory, evaluation of the mechanisms, pilot studies to contrast the effects of context or to test different modalities of interventions (in terms of feasibility, recruitment, inclusion, participation, etc.)?**
	Session chairs **Lehana Thabane, Mc Master University and Jeanine Pommier, EHESP**
15	**Introduction.** Lehana Thabane, Mc Master University
10	**Pilot study or evaluability assessment?,** Louise Potvin, Université de Montréal
10	**Exploring intervention mechanisms before piloting a smoking cessation prevention program : the RESIST study,** Laetitia Minary, Université de Montréal and Joëlle Kivits, Université de Lorraine
10	**Pilot study and process evaluation: a happy marriage?**, Kareen Nour, Ecole de Santé Publique de l'Université de Montréal
40	***Discussion***
20	***Questions of the public***
15	**Conclusion/Overview. Jeanine Pommier,** Ecole des Hautes Etudes en Santé Publique
15..30 - 16.00	***Coffee break***
16.00 - 18.00	**Session 3 : Which methodologies might be combined in process evaluation (qualitative, quantitative, mixed-method, realist approach), and how? What is the temporality of process evaluation? Might process evaluation be nested in a trial or is it an alternative design? Which methods can be mobilized? How can and should process evaluation inform intervention adaptation in PHIR? How does a systems perspective shape methods used for process evaluation?**
	Session chairs **Nancy Edwards, University of Ottawa and François Dabis, Inserm UMR 1219, Université de Bordeaux**
15	**Introduction.** Nancy Edwards, University of Ottawa
10	**Point of view (Process evaluation-part of an RCT, an alternative design or both?),** Rona Campbell, University of Bristol
10	**Point of view (Realist randomized controlled trials),** Christopher Bonel, LSHTM
10	**Mixed methods' contribution to process evaluation of population health interventions,** Marie-Renée Guével, Ecole des Hautes Etudes en Santé Publique
40	***Discussion***
20	***Questions of the public***
15	**Conclusion/Overview.** François Dabis, Inserm UMR 1219, Université de Bordeaux
18.00 - 18.15	**Conclusion of the day.** **Geneviève Chêne,** Director of Aviesan ITMO Public Health, Director of the Public Health Research Institute (IReSP)

SPEAKERS

Geneviève Chêne, Director of Aviesan ITMO Public Health, Director of the Public Health Research Institute (IReSP)

Geneviève Chêne, MD PhD, is professor in clinical epidemiology and public health at Bordeaux University since 1999.

As a researcher, between 2003 and 2015, Geneviève has led an Inserm research team on clinical research, than on "HIV infections and associated morbidity" and a Clinical Trials Unit as a platform of excellence for national and international studies in HIV/AIDS sponsored by the National Agency for Research on AIDS and viral hepatitis (ANRS). Most cited works showed how the efficacy and safety of antiretroviral therapy can be durable over the long term or contributed to the development of epidemiological and statistical innovations needed to conceive or analyse clinical trials or large cohorts. Since 2010, as co- principal investigator of Memento, a national cohort of 2,300 participants recruited in memory clinics, she is actively transferring know-how in other areas than HIV (Inserm Unit 1219).

Since 2014, as head of an international platform for international clinical trials ("EUCLID"), funded by the investments for the future through F-CRIN infrastructure, she is involved in the support and coordination of trials for the evaluation of innovative strategies in the field of vaccines (HIV, viral hepatitis, Ebola virus), neuro-degenerative diseases, obesity/metabolism or medical devices. This platform is one component of the Clinical Epidemiology Center (CIC-EC7, Inserm, CHU de Bordeaux) that she is leading since 2008. At Bordeaux University hospital (CHU Bordeaux), she is also head of the public health department since 2011.

As a professor of public health, Geneviève Chêne teaches clinical epidemiology at the Bordeaux School of Public Health (Univ Bordeaux) where she initiated an international distant-learning program in epidemiology and clinical research (more than 4500 students since 2001).

In 2013, she was nominated in the "Comité des Sages" by the Prime Minister for the preparation of the National Strategy for Health. Geneviève Chêne is currently the Director of the Public Health Multi-Organization Thematic Institute of the French National Alliance for Life Sciences and Health (Aviesan and Inserm, Paris) and, since 2013, serves as vice-chair of the scientific evaluation committee of the National Program for Clinical Research (PHRC).

Her main areas of interest are: Translational epidemiology and public health, clinical epidemiology, e-learning, infectious diseases, Alzheimer and neuro-degenerative diseases.

Christine Chomienne, Director for Research and Innovation, National Cancer Institute (INCa)

Christine Chomienne is the Director of Research and Innovation Programmes at the French National Cancer Institute (INCa) and Director of the Cancer Multi-organization thematic institutes of the French National Alliance for Life Sciences and Health (ITMO Cancer - Aviesan).

Christine Chomienne is Professor of Cellular Biology at the University Paris Diderot of Paris, France and Head of the Cell Biology Department at the Hôpital Saint Louis, Paris. She is also Director of the University Inserm Research Laboratory at the Institut Universitaire d'Hématologie. She qualified in medicine at the Université Paris Diderot and received certification for specialized training in Hematology in 1983. She obtained her PhD in 1989.

Her main research interest is in myeloid malignancies and the analysis of myeloid signalling pathways for the identification of novel therapeutic targets and strategies. She has been a key investigator in targeted therapy especially differentiation therapy, apoptosis and immunotherapy. Dr Chomienne was a pioneer researcher in differentiation therapy and translated ATRA therapy in APL in 1987. She has since devoted her time in translational research in myeloid malignancies and coordinating internal conferences and networks for dissemination and training in novel concepts and technologies in France and Europe. She is committed to education nationally and coordinator of different Courses and Masters at the University Paris Diderot. She established the Institute of PhD Schools at the University Paris Diderot. She was President of the European Hematology Association from 2013 to 2015 and is currently the immediate past-president until June 2017. Dr. Chomienne is author or co-author of more than 250 peer-reviewed publications and has received several scientific awards.

François Alla, Deputy-Director of IReSP

François Alla earned a medical doctor degree and a PhD in epidemiology. He is a professor of public health at the Université de Lorraine, where he is the director of the School of public health and the head of the "Evaluation of complex interventions" research team (EA 4360 APEMAC).

He authored or co-authored more than 200 articles, books, book chapters and reports. His research fields include clinical epidemiology, evaluation of public health interventions and research methods.

He is also the Deputy-director of the French institute for public health research (IReSP), and the Editor-in-Chief of Santé Publique, the peer-reviewed journal of the French Society of Public Health.

Pierre Arwidson, Santé Publique France, St Maurice

Dr Pierre Arwidson has studied medicine at the University François Rabelais in Tours. He wrote a thesis on problem-based learning applied to the training of medical students after studying educational innovations implemented at the University of Mc Masters and the South Illinois School of Medicine(Pr Howard Barrows) and the University of Health sciences in Linnköping (Pr Torsten Denneberg). Then appointed as an educational advisor at the Faculty of Medicine of Tours, he designed and implemented learning modules built around simulated patients(especially to teaching in ENT, Pr Beutter).

At the request of Pr Jacques Drücker and to address the HIV epidemic in the early 1990s, he has shifted to health education. In particular, he created and led a departmental network of AIDS prevention for 5 years. He gradually approached other subjects such as addiction prevention within the Departmental Committee for health education in Indre-et-Loire and the prevention of cardiovascular diseases in the Regional Heart Foundation, both chaired by Pr Mireille Brochier.

At the request of Dr François Baudier, he joined the French Council for Health Education in 1997 as deputy of Christiane Dressen, head of the survey and evaluation department. After being Director of Scientific Affairs of the National Institute of prevention and health education from 2002 to 2015, he is currently deputy director of prevention and health promotion in the newly created Public Health France.

He was vice president of the International Union for education and health promotion from 2004 to 2010. He represents Public Health France in the Commission for evaluation, strategy and prospective at the High Council for Public Health. He is an active member of the European Society for Prevention Research. He is a member of UNODC's international informal scientific network.

Pierre BLAISE, ARS Pays-de-la-Loire, Nantes

Pierre Blaise MD, MPH, PhD is a medical doctor, public health specialist. He worked in Africa from 1988 to 1995 as medical coordinator of a district health services development program led by the NGO Médecins du Monde in Guinea and as government medical officer and researcher for a EU district health services management research project in Zimbabwe. As international consultant from 1996 to 98 he conducted several consultancies on health services evaluation and planning. He joined the department of public health of the Institute of Tropical Medicine, Antwerp, Belgium in 1998. His research focuses on quality management in health services. His PhD focuses on the challenge of change in public health services in Africa. Evaluating quality management interventions, he applied the realistic evaluation method, for the first time in this field.

He joined the Health and Social Affairs Regional Direction (DRASS) in France, Pays de la Loire, as public health medical inspector in 2007. Appointed director of the Regional Health Project of the newly created Regional Health Agency (ARS) in april 2010, he wrote and managed the first regional health project running from 2012 to 2016. Today, together with the regional health partners he prepares the second generation of Regional Health Project, establishing the 10 years strategic orientations and operational objectives for the agency.

Emmanuel BONNET, Institut de Recherche pour le Développement, Paris

Emmanuel is a health geographer, a researcher at IRD (French National Institute for Sustainable Development) with international experience in spatial surveillance, GIS methodologies and spatial analysis. He is an expert in assessment of vulnerable populations especially in Africa. He is currently involved in dengue research intervention project in Burkina Faso with an international team. He proposes an assessment of intervention with a spatial focus using geographical methodologies.

Chris Bonell, London School of Hygiene and Tropical Medicine, London

Chris Bonell is Professor of Public Health Sociology and Head of the Department of Social & Environmental Health Research at the London School of Hygiene and Tropical Medicine. His main interests are in the evaluation of complex interventions and the social determinants of adolescent health and how to address these.

Isabelle Boutron, CRESS-UMR 1153, Université Paris-Descartes, Paris

Pr. Isabelle Boutron is Professor of Epidemiology at the University Paris Descartes and a researcher at the INSERM Research Center of Epidemiology and Statistics Sorbonne Paris Cité, U1153. She is an expert in the methodological characterization and evaluation of non-pharmacologic treatments, risk of bias in trials, reporting bias and the problems of transparency of research. She has a strong focus on the internal validity of non-pharmacologic trials, especially around blinding. She led with Pr Ravaud the development of international recommendations to improve the transparency of trials assessing nonpharmacologic treatments, the CONSORT non-pharmacologic interventions extension. She is also co-convenor of the Bias Method Group of the Cochrane Collaboration, and deputy director of the French EQUATOR (Enhancing the QUAlity and Transparency Of health Research) center.

Linda Cambon, Ecole des Hautes Etudes en Santé Publique, Paris

Linda Cambon, PhD in Public Health, is a professor at the Ecole des hautes études en santé publique (French national School of public health). She holds a Chair of Research in Cancer Prevention. Her researches are focused on the evaluation of complex prevention interventions, exploring their conditions of efficacy and transferability. She has notably designed the ASTAIRE-tool to assess the transferability of local health promotion interventions. She also works on the knowledge transfer strategies to bring the gap between researchers, practitioners and decision-makers. She previously had responsibilities in health ministry - as Minister advisor in charge of prevention policies and child protection, and as public health director in a Regional Agency of Health. She also led public health non-benefit organizations. She is a member of the French Society of Public Health (SFSP) board and Vice-President of the International Union of Promotion and Health Education (IUHPE).

Rona Campbell, University of Bristol, Bristol

Rona is Professor of Public Health Research, leads the Centre for Public Health Research within the School of Social and Community Medicine at the University of Bristol. She is the Bristol-based Co Director of DECIPHer and the Bristol lead for the National Institute of Health Research's School for Public Health Research (NIHR SPHR) and will shortly become Deputy Director of the national school. Rona leads programmes of research concerned with multiple risk behaviour in adolescence and health promotion in schools. She is currently involved in a number of RCTs and systematic reviews all seeking evidence for the best ways to improve the health and well-being of children and young people. Rona has a strong interest in methodological research, in particular, how to use qualitative methods alongside quantitative approaches, and how to make better use of social and behavioural theory in public health research. With Laurence Moore of the University Glasgow she founded DECIPHer Impact, a not-for-profit company dedicated to licensing and supporting the dissemination of evidenced-based, public health improvement interventions.

François Dabis, Inserm UMR 1219, Université de Bordeaux, ISPED, Bordeaux

François Dabis is a medical doctor, Professor of Epidemiology at the School of Public Health (ISPED) of the University of Bordeaux, France. He was leading from 2001 to 2015 the "HIV, cancer and global health" research team and is now a member of the "Infectious Diseases Epidemiology" Team and of the "Morbidity and Public Health HIV Hepatitis" Team of the INSERM 1219 Bordeaux Population Health Research Centre at ISPED.

Dr. Dabis has 30 years of experience in research on HIV epidemiology and global health. His scientific interest is on the public health challenges of HIV prevention and care: prevention of mother-to-child transmission for many years and now universal test and treat in Africa and in France, prognosis of antiretroviral-treated adults in West Africa and France and more generally implementation science.

François Dabis was the Chair of the Coordinated Action n°12 of the French Agency for Research on HIV/AIDS and Viral Hepatitis (ANRS) in charge of the scientific agenda of the Agency in lower-income countries from 2002 to 2015. He has been involved in international guidelines development for the World Health Organization, UNAIDS and in France for the past ten years.

François Dabis was the Chair of the French Institut national de la veille Sanitaire (InVS) from 2003 to 2012.

He has published more than 670 papers and two leading textbooks in Field Epidemiology.

Nancy Edwards, RN, PhD, Fellow of the Canadian Academy of Health Sciences, Ottawa

Nancy Edwards is a Distinguished Professor, University of Ottawa, and Full Professor in the School of Nursing. She completed an eight-year term as Scientific Director, Institute of Population and Public Health, Canadian Institutes of Health Research in July, 2016. Dr. Edwards obtained her undergraduate nursing degree from the University of Windsor and completed graduate studies in epidemiology at McMaster University and McGill University. She is a fellow of the Canadian Academy of Health Sciences.

Dr. Edwards' clinical and research interests are in the fields of public and population health. She has conducted health services, policy and clinical research both nationally and internationally. Her research has informed the design and evaluation of complex multi-level and multi-strategy community health programs. Her work in global health has spanned four continents where she has led both development-oriented and research-focused projects.

Christine Ferron, National Federation for Health Education and Promotion, Paris

Currently General Delegate of the National Federation for Health Education and Promotion (FNES) in France, Christine Ferron has previously held management responsibilities in several public and private organizations : the Regional Authority for Health Education and Promotion in Brittany, the Foundation of France, the National Institute for Prevention and Health Education, the French Committee for Health Education...

Holding a PhD in developmental psychology, she also worked as a researcher in several teams in France and abroad : the Center for Preventive Medicine of Nancy, the Psychology Department and the Department of Psychiatry and Behavioral Sciences of Northwestern University in Evanston-Chicago (as part of a Fulbright Scholarship), the Division of General Pediatrics and Adolescent Health of the University of Minnesota in Minneapolis, the University Institute of Social and Preventive Medicine in Lausanne... As an associate professor at the School of Higher Studies in Public Health (EHESP), and a health promotion specialist, she provides teaching sessions as part of initial or professional training courses.

Marie-Renée Guével, Ecole des Hautes Etudes en Santé Publique, Rennes

Lecturer in Education at the Social sciences department of the Ecole des Hautes Etudes en Santé Publique, she is part of the Centre de Recherche sur l'Action Politique en Europe (CRAPE-ARENES UMR 6051). After an engineering degree in agrifoodbusiness and a master's degree in school health education, she completed a PhD focusing on factors influencing the implementation of a health promotion approach in French primary schools. Based on an intervention research implemented in six regions, she has developed a research interest for the use of mixed methods in public health research, especially, through the evaluation of school health promotion projects. She is currently working on disability issues within the workplace and coordinating a research programme using both qualitative and quantitative approaches.

Nadir Kellou, Collège Universitaire de médecine générale, Faculté de médecin de Lyon-est, Université Claude Bernard Lyon 1, Lyon

Dr. Nadir Kellou is a general practitioner and a teaching fellow at the Faculty of Medicine Lyon-Est University Claude Bernard Lyon 1 (UCBL1) in France. Dr. Kellou received his General Practice Medicine Doctorate (MD) in 2008, his Public Health PhD degree in 2013 and his medical pedagogy degree in 2014 at the UCBL1. His research interests include physical activity and health behaviours. During his PhD, using a socioecological approach and based on a literature review, he participated to the publication of an article evaluating the effectiveness of interventions promoting physical activity for preventing unhealthy weight in children. The conclusion of this article highlights the effectiveness of intervention programs targeting all the components of the socioecological approach.

Joëlle Kivits, Université de Lorraine, Nancy

Joëlle Kivits is lecturer in sociology at the School of Public Health, University of Lorraine (Nancy, France). She is a member of the interdisciplinary research team, APEMAC. Her research works concern health education, and information and communication in public health. She also contributes to the development of innovative frameworks for evaluating complex interventions in health promotion. She teaches qualitative research and sociology of health and illness. She is associate editor to the journal « Santé publique ».

Anthony Lacouture, Ecole des Hautes Etudes en Santé Publique, Rennes et Université de Montréal, Montréal

After getting his master's degree in Public health at ISPED Bordeaux School of Public Health in 2011, Anthony Lacouture was in charge at the French National Institute for Prevention and Health Education (Inpes) of a program evaluation and international expertise on the accessibility of information in prevention and health promotion to disabled people, and especially deaf or visually impaired people. Following that, he has been working as research assistant at the Inpes Chair in Health Promotion at EHESP on the realist approach in program evaluation. An article has been recently published in the Implementation Science review. In September 2013, he started doing his PhD in Public Health and Political Science in partnership with the University of Montreal School of Public Health (ESPUM) and the University of Rennes 1 - EHESP School of Public Health. His work is part of the RICAP research program "Research and Intervention: Collaboration between researchers and local policymakers in health promotion" funded by Inpes. His research interests focus on program evaluation, integrated knowledge translation, and professional practices in health promotion.

Thierry Lang, UMR 1027, Inserm-Université Toulouse III Paul Sabatier, Toulouse

Director of the « Institut Fédératif d'Etudes et de Recherches Interdisciplinaires Santé Société » (IFERISS) http://www.iferiss.org/

Epidémiologist, Professor at Toulouse III University and Toulouse University Hospital Team « Inequalities in health, cancer and chronic diseases» from Unit 1027 INSERM –Université Paul Sabatier, Toulouse 3 (2002-2015). Member of the High Council for Public Health (HCSP), chairman of the working group on Inequalities in Health (Reports published in 2010, 2013, 2016). Responsible of the Research Master Clinical Epidemiology: http://www.biostat.envt.fr/master/

Susan Michie, University College London, London

Susan Michie is Professor of Health Psychology at University College London, UK. She studied Experimental Psychology at Oxford University, followed by Clinical Psychology at the Institute of Psychiatry, London University and a DPhil in Developmental Psychology. She is a chartered clinical and health psychologist, and elected Fellow of the Academy of Social Sciences, the US Society of Behavioral Medicine, the US Academy of Behavioral Medicine Research, the European Health Psychology Society and the British Psychological Society.

Professor Michie is Director of the Centre for Behaviour Change (http://www.ucl.ac.uk/behaviour-change) and of the Health Psychology Research Group at UCL. She leads an extensive programme of research developing the science of behaviour change interventions and applying that science to intervention development and evaluation. Areas of application focus on prevention of ill health and implementation of evidence-based practice. Methodological projects include the Wellcome Trust-funded Human Behaviour-Change Project (www.humanbehaviourchange.org) and the MRC- funded Theory and Techniques project (www.ucl.ac.uk/behaviour-change-techniques).

Personal website: www.ucl.ac.uk/health-psychology/pages/michie

Laetitia Minary, Université de Montréal, Montréal

Laetitia Minary is a researcher at the University of Lorraine in the EA 4360 APEMAC team " Chronic diseases, perceived health and adaptation process ". She is also a visiting Professor at the University of Montreal in the Department of Social and Preventive Medicine of the School of Public Health. She holds a PhD in public health and epidemiology. Her main research areas are on evaluation of complex interventions in public health, and she is specifically interested in innovative methods of evaluation. The scope of her research is smoking cessation programmes implemented in vulnerable adolescents.

Graham Moore, Cardiff University, Cardiff

Graham Moore is Deputy Director of DECIPHer (Cardiff University) and Senior Lecturer in Social Sciences and Health. He currently leads DECIPHer's programme of research and teaching around methodology for evaluating complex interventions. He was lead author of the MRC guidance for process evaluation of complex interventions, which emphasises the need to understand implementation and causal mechanisms, as well as how interventions and their contexts interact, in order to make sense of outcomes evaluation data. Substantively, his interests are in i) tobacco control policy and the denormalisation of smoking, and ii) the role of schools and school based intervention in increasing or reducing socioeconomic inequalities in young people's health and health behaviours.

Grégory Ninot, Université de Montpellier, Montpellier

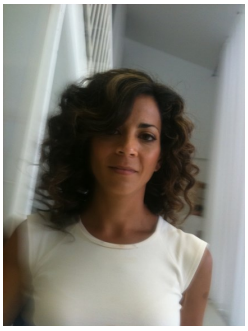
My research activities revolve around the verification of the efficacy, the safety, and the cost-effectiveness of non-pharmacological interventions (NPIs), as well as their impact on patient satisfaction. NPIs consist in exercise methods, physiotherapy programs, nutrition interventions, disease management educational programs and psychotherapy methods.

Targeted populations are patients with chronic diseases (e.g., COPD, cancer) or persons at major risk for diseases (e.g., risk of falling). I work on interventional studies and clinical trials, as well as meta-analyses and systematic reviews on this topic. I am also involved in the validation of self-reported questionnaires using psychometric standards. My research encompasses concepts such as chronic disease acceptance, quality of life, adherence, anxiety, depression, fatigue, self-esteem and their impact on health behaviors over time. The overarching purpose of these activities is to contribute to the improvement of care and health prevention solutions.

Professor, University of Montpellier, France

Executive Director CEPS Platform www.CEPSplatform.eu
Integrated Cancer Research Site (SIRIC) www.montpellier-cancer.com

Kareen NOUR, Ecole de Santé Publique de l'Université de Montréal, Montréal



Graduated from University of Montreal with a PhD in public health, Ms. Kareen Nour is researcher at the Public Health Department in Montérégie. She is also a clinical professor at the Department of social and preventive medicine at the School of Public Health at the University of Montreal, an associate professor at the University of Sherbrooke and a member of various research groups in Quebec (Canada). Over the years, she has developed a particular expertise in evaluative research looking at public health programs and front-line services. Her main research projects explore the implementation of programs in natural settings that involve cross-sectorial partners. For example, some of her work evaluates the implantation and the effects of *health impact assessment* (HIA), others explore the trajectories of young receiving mental health services or others analyze the implementation of a procedure for suicide prevention. Combining qualitative and quantitative approaches, she obtained different research grants as a principal investigator or co-investigator. She also supervises internship, master's degree or doctorate students. She has published numerous articles and has made several scientific presentations. Finally, she is an evaluator for various private and public funding agencies.

Jeanine Pommier, Ecole des Hautes Etudes en Santé Publique, Rennes

Jeanine Pommier, MD, PhD, is professor at the French School of Public Health and deputy of the department of Social and Human sciences. She is a researcher at the Research Center on Policy action in Europe.

She is currently developing realist evaluation projects and assessing the integration of public policies to reduce health disparities in the French public health system. Furthermore, she is developing a research project in Knowledge transfer in order to develop better informed public health policy in France.

Louise Potvin, Université de Montréal, Montréal

Louise Potvin is currently professor at the Department of Social and Preventive Medicine, School of Public Health, Université de Montreal and Research at the Institut de recherche en santé publique, Université de Montréal and at the Centre Léa-Roback sur les inégalités sociales de santé de Montréal. She holds the Canada Research Chair in Community Approaches and Health Inequalities. Her main research interests are Population Health Intervention Research and the role of social environments in the local production of health and health equity. In addition to having edited and co-edited 8 books, she has published more than 250 peer-reviewed papers, book chapters, editorials and comments. She is a Fellow of the Canadian Academy of Health Sciences and the Editor in Chief of the Canadian Journal of Public Health.

Lehana Thabane, Mc Master University, Hamilton

Dr Lehana Thabane is a Professor of Biostatistics and Associate Chair of the Department of Clinical Epidemiology and Biostatistics (CE&B), Associate member of the Departments of Pediatrics and Anesthesia in the Faculty of Health Sciences (FHS) at McMaster University (Hamilton, Ontario, Canada). He is also the Director of Biostatistics at St Joseph's Healthcare—Hamilton (Ontario, Canada).

He has an excellent track record as a lead /senior biostatistician for over 100 externally funded studies. Highly sought out speaker at international conferences, his research interests include clinical trials, primary care, evidence-based medicine, mentorship. He has co-authored over 470 publications in peer-reviewed journals and over 500 abstracts presented at national and international meetings.

A winner of the CE&B Excellence Award in Teaching for 2004-2006; the FHS Excellence in Graduate Supervision Award for 2012; and the McMaster President's Excellence in Graduate Supervision Award for 2016, Dr Thabane has extensive experience as an educator and mentor. To date he has mentored over 100 MSc, PhD and Postdoc trainees. He is the clinical trials mentor for the Canadian Institutes of Health Research.

ABSTRACTS OF THE PRESENTATIONS

Session 1 : The place of theory into process evaluation

Behaviour change techniques and their mechanisms of action. Susan Michie, University College London

One of the objectives of a process evaluation is to understand the mechanisms by which an intervention has had its effect, in order to develop more effective interventions. In the case of complex interventions, such as most or those aimed at changing behaviour, it is necessary also to understand which components within a complex intervention are linked with which mechanisms. This talk will present the findings of a research programme linking behaviour change techniques to their mechanisms of action. These can be used to (i) support the development of theory-based interventions and (ii) enable the theoretical understanding of empirical evaluations of interventions not explicitly based on theory.

A Need of Integrative and Comprehensive Health Intervention Ontology for Intervention Research. Grégory Ninot, Université de Montpellier

Integrated and Comprehensive Health Interventions (ICHIs), also called Non Pharmacological Interventions (NPIs), have become essential solutions to improve health, quality of life and, often, life expectancy. Recent studies have also highlighted the positive social and economic impact. "ICHIs are non-invasive methods of care (programs, products or services) whose efficacy in improving the health and quality of life of human beings has been proven. Their effects on health and quality of life markers are observable (with measured risks and benefits beyond mere user opinions) and can be linked to identified biological and/or psychological processes. They can also have a positive impact on health behaviours and socio-economic indicators" (CEPS Platform, 2016). Many authors (e.g., Ioannidis, 2015) and health authorities argue that what has been brought forth is merely proof of concept (e.g., French Health Authority, 2011). They note the lack of a consensus paradigm of validation and surveillance, such as the standards in drug development. They highlight the methodology problems due to rapid obsolescence of ICHI using a digital solution (e.g., Mobile Apps). As a result, policy makers and health decision-makers remain skeptical of the impact of ICHIs. These key players are encouraging innovators to come forward with additional evidence for the efficacy and the cost/effectiveness of ICHIs in order to improve their visibility, and, ultimately, to garner more substantive private and public financial support for them. ICHIs need to be compared and optimized, as well as targeted to the right health problem at the right time. The first step will be to develop a collaborative top-down strategy to identify and classify these health interventions. The communication presents our strategy supported by French State, Occitanie Region, and Montpellier Metropole.

Socioecological theory-based interventions: How to evaluate the effectiveness of their mechanisms? Nadir Kellou, Collège Universitaire de médecine générale, Faculté de médecin de Lyon-est, Université Claude Bernard Lyon 1

The socioecological approach is being increasingly used in intervention studies that aim to promote healthier behaviours. From physical activity promotion to obesity prevention, programs based on this approach have achieved results that seem promising. Yet the question about the effectiveness of their mechanisms is still open and the literature dealing with this question is relatively scarce. One of the possibilities to assess the effectiveness of their mechanisms could be through an evaluation of their internal and external validity. The internal validity would be assessed by controlling, among of other things, that: The intervention has a positive effect on a health outcome; the methodology is not biased; the intervention is equitable, the intervention effectiveness is theory-driven; and the health effects produced by the intervention are sustainable. The external validity would be assessed by controlling, among of other things, that: The individual participation rate is high; the program is correctly implemented; the program is cost-effective and the intervention effects are both reproductible and transferable. Thus the process evaluation of socioecological theory-based interventions would be feasible but would require to be anticipated during the elaboration of the study protocol.

Session 2 : The place of pilot studies in process evaluation

Pilot study or evaluability assessment? Louise Potvin, Université de Montréal

There are great areas of overlap between the field of population health intervention research and that of evaluation. This presentation will examine how the concept of evaluability assessment developed by evaluators in over the past 30 years can inform the role of pilot studies for population health intervention research

Exploring intervention mechanisms before piloting a smoking cessation prevention program : the RESIST study. Laetitia Minary, Université de Montréal and Joëlle Kivits, Université de Lorraine

Most smoking prevention programs targeting adolescent population focus on preventing smoking initiation with limited attention to smoking cessation. However in France, more than 30% of 17 year olds are smokers, this proportion rising up to 50% for the apprentices, a particularly vulnerable population regarding health. The TABADO program targeting young vocational trainees has recently shown its effectiveness. It included an informative meeting about smoking for all students, and for smokers who wished to participate, an enhanced program (EP) combining: motivational counseling, medication and cognitive behavioral therapy sessions provided by smoking cessation specialist. In the RESIST project, we suggest to optimize this program by adding a strategy based on the influence of social networks that may optimize the effects of intervention by favoring the participation of youth in the intervention but also promoting their adhesion and maintaining health behavior change.
Prior to evaluate the efficacy of the optimized intervention, an exploration phase of the mechanisms of the initial TABADO intervention, including a social network analysis, will help to develop the “social support” strategy that could optimize the results of the TABADO intervention. A pilot study of the RESIST intervention (TABADO + “social support strategy”) and research will then be carried out to determine adaptations to be made to the initial intervention. The exploration phase of the project is here presented. We will expose the global approach to analyze implementation, psychological, sociological and epidemiological mechanisms influencing the effect of the TABADO intervention.

Pilot study and process evaluation: a happy marriage? Kareen Nour, Ecole de Santé Publique de l'Université de Montréal

Pilot studies and process evaluation are two domains in the field of interventional research and public health that are growing in interest. While pitot studies are crucial for research to test a large study or for intervention to explore program strategies, process evaluation give a lot of insight about how an intervention is implemented and why it might have potential impact on health outcomes health. An increasing number of studies combined pilot studies and process evaluation both. But how to do it? Why choosing such research design? What are the benefits or limits of such researches design? What objectives that can be suitable by doing a process evaluation in a pilot study? What methodology should be used? All those questions will be addressed in this workshop with example of evaluation done in this context.

Session 3 : The combination of methods in process evaluation

Point of view (Process evaluation-part of an RCT, an alternative design or both?). Rona Campbell, University of Bristol

Drawing on 20 years' experience of undertaking qualitative research and process evaluation within pilot and definitive RCTs of complex public health interventions, this presentation will offer some reflections on the question of whether process evaluation should be nested within a trial or whether it comprises an alternative design. It will suggest that these are not either or options but that process evaluation can be considered both as a separate entity and one that should always be embedded with RCTs of complex interventions. Achieving this, however, is problematic, and the implications in terms of incommensurability and temporality will be considered alongside the more practical issues of ways of managing such trials, and the different skills sets required to operationalise such an approach. The presentation will also consider recent argument that there is an ethical imperative to having process evaluations accompany RCTs of public health interventions but that they need to be clearly and formally recognised as a distinct entity. The presentation will conclude by observing that single trials rarely provide all the evidence required and that it is important to consider any process evaluation within one trial in the context of other relevant evidence including evidence syntheses.

Point of view (Realist randomized controlled trials). Christopher Bonel, LSHTM

Evaluations needs to examine how and for whom not merely whether interventions are effective. Realist evaluation presents a useful framework for this, proposing that to understand interventions we must understand how they might lead to social mechanisms which interact with context to produce outcomes. Evaluators must not merely understand effectiveness in terms of statistical associations between intervention exposure and outcomes but must instead develop theories and hypotheses about mechanisms which they can then test with empirical data. Whereas realist evaluation has traditionally argued that randomised controlled trials are inappropriate for realist evaluation because of their positivist assumptions, we argue that randomised trials are not necessarily positivist and can in some but not all circumstances should be useful tools for assessing the validity of theories about context -mechanism-outcome configurations.

Mixed methods' contribution to process evaluation of population health interventions. Marie-Renée Guével, Ecole des Hautes Etudes en Santé Publique

By combining quantitative and qualitative methods, mixed methods aim to produce a more complete picture of the phenomena being studied. When it comes to evaluation of population health interventions, mixed methods could be an interesting methodological option to cover both effectiveness and process evaluation. This presentation will illustrate contributions and challenges of the use of mixed methods in population health intervention research.

PRESENTATION OF THE HOSTING INSTITUTIONS

The French National Alliance for Health and Life Sciences

aviesan
alliance nationale
pour les sciences de la vie et de la santé

The French National Alliance for Health and Life Sciences (Aviesan) was reated around common goals with this very purpose in mind, bringing together CEA, CNRS, the French Conference of University Hospital CEOs, the French Conference of University Presidents, Inra, INRIA, Inserm, the Institut Pasteur and IRD.

An innovative organization

Aviesan was set up to enhance the French research capacities, by:

- Developing a high level continuum in every branch of health and life science research, ranging from basic research to its application.
- Strengthening partnerships between universities and research organizations, while ensuring that programmatic themes and infrastructure projects are coordinated and consistent nationwide.
- Disseminating knowledge and promoting research findings, from an industrial, clinical and social perspective.
- Defining common views, especially in relation to European research and international cooperation.
- Harmonizing and simplifying administrative processes for laboratories.

Aviesan is organized into thematic multi-organization institutions (ITMOs), working towards major objectives:

- To provide national strategic analyses and new programming capacities by bringing together the best scientists from all sources, and oversight of the scientific communities concerned (scientific coordination).
- In keeping with the strategic analyses, to foster the development of large research centres and programmes and compile biological and data processing resources through initiatives that are decided jointly by national research organisations and universities (operational coordination).

The Health and Life Sciences Research's Coordination Council, including the Chairmen and/or CEOs of Aviesan partners and the Directors of the ITMOs, takes on the operational coordination of the different research's operators and the representation of them close to the governmental authorities, European organizations or industries.

All research fields covered

The sector of life sciences and health was broken down into 9 major themes:

- Cancer
- Cell biology, development and evolution
- Genetics, genomics and bioinformatics
- Health technologies
- Immunology, inflammation, infectiology and microbiology
- Molecular and structural bases of living organisms
- Neurosciences, cognitive sciences, neurology
- and psychiatry
- Pathophysiology, metabolism and nutrition
- Public health

Regarding research and innovation, Alliance Aviesan contributes to the implementation of several national plans or actions such as the neurodegenerative disorders plan, the antibioresistance plan, the next-generation sequencing mission, the disabilities plan, the cancer plan,…

The French Research Agency ANRS

The French Research Agency ANRS (France Recherche Nord&Sud Sida-HIV Hépatites) was set up in 1988. It brings together researchers from different fields and institutions in the developed world and resource-limited countries to study scientific questions. In 2012, the ANRS became an autonomous agency of Inserm (French National Institute of Health and Medical Research).

The ANRS has an annual budget of about 48 million euros, most of which is provided by the French state (ministries of research and health). The ANRS funds research projects approved by international expert committees. It oversees projects from conception to completion and ensures that the results are used for the benefit of the populations concerned. The ANRS aims to promote excellence and to expedite data collection.

The ANRS reviews its research priorities in light of advances in understanding and the needs of affected populations. All aspects of HIV/AIDS and hepatitis are taken into account in seeking to provide coherent answers. By mobilizing teams working on fundamental, clinical, economic, public health, and human and social sciences research, the ANRS provides comprehensive responses to scientific questions in prevention, screening, and healthcare.

Together with resource-limited countries, the ANRS sets up research programs with a long-term outlook that take account of national public health priorities. In conducting their work, researchers from developed countries and resource-limited settings undertake to abide by the principles of the ANRS code of ethics.

The French National Cancer Institute

The French National Cancer Institute (INCa) is the preeminent health and science agency in charge of cancer control in France. It reports to the ministries for Health and for Research.

The Institute provides an integrated approach encompassing all cancer-control dimensions (health, scientific, social and economic) and areas of intervention (prevention, screening, care and research) for the benefit of patients and their relatives.

To catalyze progress, the INCa acts as an interface with patients, their friends and families, the healthcare system users, general public, healthcare professionals, researchers, experts and decision-makers.

The INCa missions are :

- To provide an integrated approach to cancer control ;
- To support innovation ;
- To produce evidence-based guidelines for decision-makers and professionals ;
- To coordinate regional oncology networks ;
- To analyze data to guide action more effectively ;
- To disseminate knowledge about cancer.

Inserm

Founded in 1964, the French National Institute of Health and Medical Research (Inserm) is a **public scientific and technological institute** which operates under the joint authority of the French Ministry of Health and French Ministry of Research.

As the **only French public research institute to focus entirely on human health**, in 2008 Inserm took on the responsibility for the strategic, scientific and operational coordination of biomedical research. This key role as coordinator comes naturally to Inserm thanks to the scientific quality of its teams and its ability to conduct translational research, from the laboratory to the patient's bed.
The decree adopted in March 2009 will enable Inserm to perform its research missions in the face of the new scientific, health and economic challenges of the 21st century. **Scientific monitoring and expertise** are now part of the Institute's official missions.

In early 2008, **9 thematic institutes** were created in the light of this new coordination role, in association with Inserm. The aforementioned decree secures a long-term future for them and clearly defines their remit, an inventory of French research in their field, how this research is to be managed and their objectives.

From the outset, Institute has forged close partnerships with the other public and private research establishments as well as hospitals to fulfil its missions. **80% of Inserm's 281 research units** are currently set up in university hospitals or cancer research centers. The research campuses of the French National Center for Scientific Research (CNRS), along with the Pasteur and Curie Institutes, also house Inserm research divisions. With the law on the independence of universities placing them at the heart of the research policy, they will also be a key partner of Inserm.

In April 2009, national coordination was strengthened by the Alliance nationale pour les sciences de la vie et de la santé (French National Alliance for Life and Health Sciences), which Inserm co-founded with other research institutes and the Conférence des présidents d'université (Association of University Presidents). To extend the strategic and programmatic coordination of research to all life and health sciences, the Alliance relies on 10 multi-body thematic institutes jointly overseen by two research institutes (Inserm, CNRS, French Atomatic Energy Commission/CEA or French National Institute for Agricultural Research/Inra), depending on the research field.

Lastly, Inserm plays a leading role in creating the European Research Area and boosts its standing abroad through close partnerships (teams and partner laboratories abroad).

The French Public Health Research Institute

The French Public Health Research Institute is established since 2004 under a collaborative agreement and remodelled itself as a Groupement d'Interêt Scientifique (GIS) in 2007. For financial management, the group is assisted by Inserm.

The objective of the IReSP is to develop and promote research in Public Health in France, with the help of a partnership of 23 members (Ministries, Research Institutions, Health Agencies and Social Protection Organisations).

Research fields of interest

- Health services research
- Health public policies
- Interactions of health determinants

The main actions of the IReSP

- Making an inventory of public health research teams and cohort studies in France
- Funding public health research projects
- Supporting new emerging public health research groups
- Making available resources and tools for public health research
- Supporting events promoting public health research :
- Quarterly journal: "Public Health Issues"

The Coordinated Action for Intervention Research

The national initiative of Coordinated Action for Intervention Research is carried out by the ITMO Public Health, Cancer, and Immunology, inflammation, infectious diseases and microbiology (I3M) of Aviesan Alliance and the IReSP. The aim of the coordinated action is to gather the researchers and the stakeholders of intervention research and support in France the development of population health intervention research promoting innovation in the field and useful for public decision.

The main objectives are to promote an excellent and relevant intervention research, to foster a transversal thinking and to discuss about the existing methodologies, practices and difficulties. It also contributes to shape the field of intervention research: it supports call for research proposal, cross-disciplinary scientific working group, etc.

This Coordinated Action for Intervention Research takes the form of regular meetings (3 times a year) and workshops for three years (2015-2018).

REFERENCES

Moore GF, Audrey S, Barker M, Bond L, Bonell C, Hardeman W, et al. Process evaluation of complex interventions:MedicalResearchCouncilguidance.BMJ.2015;350:h1258.

Campbell M, Fitzpatrick R, Haines A, Kinmonth AL, Sandercock P, Spiegelhalter D, et al. Framework for design and evaluation of complex interventions to improve health. BMJ. 2000 Sep 16;321(7262):694–6.

Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M, et al. Developing and evaluating complex interventions: the new Medical Research Council guidance. BMJ. 2008;337:a1655.

Shiell A, Hawe P, Gold L. Complex interventions or complex systems? Implications for health economic evaluation.BMJ.2008Jun5;336(7656):1281–3.

Hawe P, Shiell A, Riley T. Complex interventions: how "out of control" can a randomised controlled trial be?BMJ2004;328:1561.

De Silva MJ, Breuer E, Lee L, Asher L, Chowdhary N, Lund C, Patel V. Theory of Change: a theory-driven approach to enhance the Medical Research Council's framework for complex interventions. Trials. 2014 Jul5;15:267.

Fletcher A, Jamal F, Moore G, Murphy RE, Bonell C, Realist complex intervention science: Applying realist principles across all phases of the Medical Research Council framework for developing and evaluating complex interventions. Evaluation (Lond) 2016 Jul; 22(3): 286–303.

Dalkin SM, Greenhalgh J, Jones D, Cunningham B, Lhussier M. What's in a mechanism? Development of a key concept in realist evaluation. Implement Sci. 2015 Apr 16;10:49.

Hawe P, Potvin L. What is population health intervention research? Can J Public Health. 2009 Jan-Feb;100(1):SupplI8-14.

Faggiano F, Allara E, Giannotta F, Molinar R, Sumnall H, et al. (2014) Europe Needs a Central, Transparent, and Evidence-Based Approval Process for Behavioural Prevention Interventions. PLoS Med 11(10): e1001740. doi:10.1371/journal.pmed.1001740.

J. David Hawkins, PhD; Sabrina Oesterle, PhD; Eric C. Brown, PhD; Robert D. Abbott, PhD; Richard F. Catalano, PhD, Youth Problem Behaviors 8 Years After Implementing the Communities That are Prevention System A Community-Randomized Trial, *JAMA Pediatr*. 2014;168(2):122-129. doi:10.1001.

Terrie E. Moffitta, Louise Arseneault, Daniel Belskya Nigel Dicksonc Robert J. Hancox, HonaLee Harrington, Renate Houts, Richie Poulton, Brent W. Roberts, Stephen Ross, Malcolm R. Sears, W. Murray Thomson and Avshalom Caspi, A gradient of childhood self-control predicts health, wealth, and public safety, PNAS, February 15, 2011, vol. 108, no. 7, 2693–2698.

Lehana Thabane, Jinhui Ma, Rong Chu, Ji Cheng, Afisi Ismaila, Lorena P Rios, Reid Robson, Marroon Thabane, Lora Giangregorio, Charles H Goldsmith, A tutorial on pilot studies: the what, why and how. BMC Medical Research Methodology 2010 10:1.

Eldridge SM, Lancaster GA, Campbell MJ, Thabane L, Hopewell S, Coleman CL, et al. (2016), Defining Feasibility and Pilot Studies in Preparation for Randomised Controlled Trials: Development of a Conceptual Framework. PLoS ONE 11(3): e0150205. doi:10.1371/journal.pone.0150205.

Amy L. Whitehead, Benjamin G.O. Sully, Michael J. Campbell, Pilot and feasibility studies: Is there a difference from each other and from a randomised controlled trial?, Contemporary Clinical Trials 38, (2014) 130-133.

Sandra M Eldridge, Claire L Chan, Michael J Campbell, Christine M Bond, Sally Hopewell, Lehana Thabane, Gillian A Lancaster on behalf of the PAFS consensus group, CONSORT 2010 statement: extension to randomised pilot and feasibility trials, *BMJ* 2016;355:i5239.

Van Belle et al.: How to develop a theory-driven evaluation design? Lessons learned from an adolescent sexual and reproductive health programme in est Africa. BMC Public Health 2010 10:741.

Mary E. Northridge and Sara S. Metcalf, Enhancing implementation science by applying best principles of systems science, Health Research Policy and Systems (2016) 14:74.

Hueiming Liu, Janini Muhunthan, Adina Hayek, Maree Hackett, Tracey-Lea Laba, David Peiris and Stephen Jan, Examining the use of process evaluations of randomised controlled trials of complex interventions addressing chronic disease in primary health care—a systematic review Protocol, Systematic Reviews (2016) 5:138.

Process evaluation of complex interventions: Medical Research Council guidance

Graham F Moore,[1] Suzanne Audrey,[2] Mary Barker,[3] Lyndal Bond,[4] Chris Bonell,[5] Wendy Hardeman,[6] Laurence Moore,[7] Alicia O'Cathain,[8] Tannaze Tinati,[3] Daniel Wight,[7] Janis Baird[3]

[1]DECIPHer UKCRC Public Health Research Centre of Excellence, School of Social Sciences, Cardiff University, Cardiff, UK
[2]DECIPHer UKCRC Public Health Research Centre of Excellence, School of Social and Community Medicine, University of Bristol, Bristol, UK
[3]MRC Lifecourse Epidemiology Unit, University of Southampton, Southampton, UK
[4]Centre of Excellence in Intervention and Prevention Science, Melbourne, VIC Australia
[5]Department of Childhood, Families and Health, Institute of Education, University of London, London, UK
[6]Primary Care Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK
[7]MRC/CSO Social and Public Health Sciences Unit, University of Glasgow, Glasgow, UK
[8]School of Health and Related Research, University of Sheffield, Sheffield, UK

Correspondence to: G F Moore MooreG@cardiff.ac.uk

Cite this as: *BMJ* 2015;350:h1258
doi:10.1136/bmj.h1258
Accepted: 13 January 2015

Process evaluation is an essential part of designing and testing complex interventions. New MRC guidance provides a framework for conducting and reporting process evaluation studies

Attempts to tackle problems such as smoking and obesity increasingly use complex interventions. These are commonly defined as interventions that comprise multiple interacting components, although additional dimensions of complexity include the difficulty of their implementation and the number of organisational levels they target.[1] Randomised controlled trials are regarded as the gold standard for establishing the effectiveness of interventions, when randomisation is feasible. However, effect sizes do not provide policy makers with information on how an intervention might be replicated in their specific context, or whether trial outcomes will be reproduced. Earlier MRC guidance for evaluating complex interventions focused on randomised trials, making no mention of process evaluation.[2] Updated guidance recognised the value of process evaluation within trials, stating that it "can be used to assess fidelity and quality of implementation, clarify causal mechanisms and identify contextual factors associated with variation in outcomes."[3] However, it did not provide guidance for carrying out process evaluation.

Developing guidance for process evaluation

In 2010, a workshop funded by the MRC Population Health Science Research Network discussed the need for guidance on process evaluation.[4] There was consensus that researchers, funders, and reviewers would benefit from guidance. A group of researchers with experience and expertise in evaluating complex interventions was assembled to produce the guidance. In line with the principles followed in developing earlier MRC guidance documents, draft guidance was produced drawing on literature reviews, process evaluation case studies, workshops, and discussions at conferences and seminars. It was then circulated to academic, policy, and practice stakeholders for comment. Around 30 stakeholders provided written comments on the draft structure, while others commented during conference workshops run throughout the development process. A full draft was recirculated for further review, before being revised and approved by key MRC funding panels.

Although the aim was to provide guidance on process evaluation of public health interventions, the guidance is highly relevant to complex intervention research in other domains, such as health services and education. The full guidance (www.populationhealthsciences.org/Process-Evaluation-Guidance.html) begins by setting out the need for process evaluation. It then presents a review of influential theories and frameworks which informed its development, before offering practical recommendations, and six detailed case studies. In this article, we provide an overview of the new framework and summarise our practical recommendations using one of the case studies as an example.

MRC process evaluation framework

The new framework builds on the process evaluation themes described in the 2008 MRC complex interventions guidance (fig 1).[3] Although the role of theory within evaluation is contested,[5 6] we concur with the position set out in the 2008 guidance, which argued that an understanding of the causal assumptions underpinning the intervention and use of evaluation to understand how interventions work in practice are vital in building an evidence base that informs policy and practice.[1] Causal assumptions may be drawn from social science theory, although complex interventions will often also be informed by other factors such as past experience or common sense. An intervention as simple as a health information leaflet, for example, may reflect an assumption that increased knowledge of health consequences will trigger behavioural change. Explicitly stating causal assumptions about how the intervention will work can allow external scrutiny of its plausibility and help evaluators decide which aspects of the intervention or its context to prioritise for investigation. Our framework also emphasises the relations between implementation, mechanisms, and context. For example, implementation of a new intervention will be

SUMMARY POINTS

- MRC guidance for developing and evaluating complex interventions recognised the importance of process evaluation within trials but did not provide guidance for its conduct
- This article presents a framework for process evaluation, building on the three themes for process evaluation described in 2008 MRC guidance (implementation, mechanisms, and context)
- It argues for a systematic approach to designing and conducting process evaluations, drawing on clear descriptions of intervention theory and identification of key process questions
- While each process evaluation will be different, the guidance facilitates planning and conducting a process evaluation

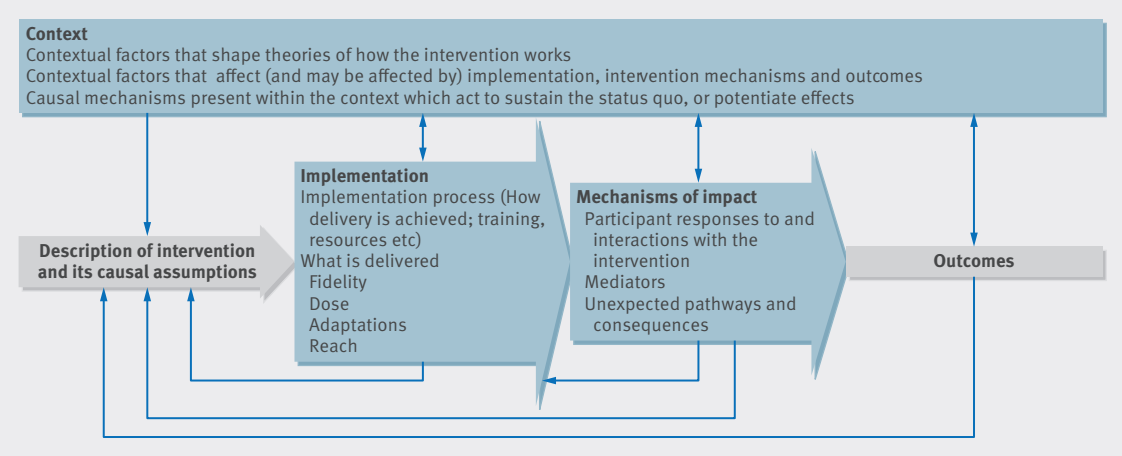


Fig 1 | **Key functions of process evaluation and relations among them (blue boxes are the key components of a process evaluation. Investigation of these components is shaped by a clear intervention description and informs interpretation of outcomes)**

affected by its existing context, but a new intervention may also in turn change aspects of the context in which it is delivered.

Implementation: what is implemented, and how?

An intervention may have limited effects either because of weaknesses in its design or because it is not properly implemented.[7] On the other hand, positive outcomes can sometimes be achieved even when an intervention was not delivered fully as intended.[8] Hence, to begin to enable conclusions about what works, process evaluation will usually aim to capture fidelity (whether the intervention was delivered as intended) and dose (the quantity of intervention implemented). Complex interventions usually undergo some tailoring when implemented in different contexts. Capturing what is delivered in practice, with close reference to the theory of the intervention, can enable evaluators to distinguish between adaptations to make the intervention fit different contexts and changes that undermine intervention fidelity.[9 10] Unresolved debates regarding adaption of interventions, and what is meant by intervention fidelity, are discussed at length in the full guidance.

In addition to what was delivered, process evaluation can usefully investigate how the intervention was delivered.[11 12] This can provide policy makers and practitioners with vital information about how the intervention might be replicated, as well as generalisable knowledge on how to implement complex interventions. Issues considered may include training and support, communication and management structures, and how these structures interact with implementers' attitudes and circumstances to shape the intervention.

Process evaluations also commonly investigate the "reach" of interventions (whether the intended audience comes into contact with the intervention, and how).[13] There is no consensus on how best to divide the study of implementation into key subcomponents (such as fidelity, dose, and reach), and it is currently not possible to adjudicate between the various frameworks that attempt to do this. These issues are discussed further in the full guidance document.

Mechanisms of impact: how does the delivered intervention produce change?

Exploring the mechanisms through which interventions bring about change is crucial to understanding both how the effects of the specific intervention occurred and how these effects might be replicated by similar future interventions.[14] Process evaluations may test hypothesised causal pathways using quantitative data as well as using qualitative methods to better understand complex pathways or to identify unexpected mechanisms.[15]

Context: how does context affect implementation and outcomes?

Context includes anything external to the intervention that may act as a barrier or facilitator to its implementation, or its effects. As described above, implementation will often vary from one context to another. However, an intervention may have different effects in different contexts even if its implementation does not vary.[16] Complex interventions work by introducing mechanisms that are sufficiently suited to their context to produce change,[17] while causes of problems targeted by interventions may differ from one context to another. Understanding context is therefore critical in interpreting the findings of a specific evaluation and generalising beyond it. Even where an intervention itself is relatively simple, its interaction with its context may still be highly complex.

Functions of process evaluation at different stages of development, evaluation, and implementation

The focus of process evaluation will vary according to the stage at which it is conducted. The MRC framework recommends a feasibility and piloting phase after an intervention has been developed.[1 3] At this stage, process evaluation can have a vital role in understanding the feasibility of the intervention and optimising its design and evaluation. However, at the next stage,

doi: 10.1136/bmj.h1258 | *BMJ* 2015;350:h1258 | the bmj

evaluating effectiveness, the emphasis of process evaluation shifts towards providing greater confidence in conclusions about effectiveness by assessing the quantity and quality of what was delivered, and assessing the generalisability of its effectiveness by understanding the role of context. Even when a process evaluation has been conducted at the feasibility stage, another will usually be needed alongside the full trial because new problems are likely to emerge when the intervention is tested in a larger more diverse sample.

Planning, designing, conducting, and reporting a process evaluation

Box 1 summarises the key recommendations of the new MRC guidance for process evaluation. Given the diversity of complex interventions, the aims and methods of process evaluations will vary, but there are common considerations when developing and planning any such evaluation. The recommendations are not intended to be prescriptive but to help researchers to make decisions. Throughout this section, we have illustrated our points using one of the six case studies included in the full guidance, the process evaluation of the Welsh national exercise referral scheme (NERS)[8 18 19]; this scheme aimed to improve physical activity through primary care referral to exercise professionals in local authority leisure centres.

Planning a process evaluation

Working with intervention developers and implementers

High quality process evaluation requires good working relationships with all stakeholders involved in intervention development or implementation. These can be difficult to establish—for example, because these stakeholders have professional or personal interests in portraying the intervention positively, or see evaluation as threatening. However, without good relationships, close observation of the intervention can be challenging. Evaluators also need to ensure that they maintain sufficient independence to observe the work of stakeholders critically. The NERS process evaluation identified serious problems with the implementation of some intervention components.[19] Evaluators needed to be close enough to the intervention to record these problems and understand why they occurred, yet sufficiently independent to report them to intervention stakeholders honestly. Transparent reporting of relationships with policy and practice stakeholders, and being mindful of how these affect the evaluation, is crucial.

One key challenge in working with intervention stakeholders is whether to communicate emerging findings. That is, should evaluators act as passive observers who feed findings back at the end of an evaluation or help to correct problems in implementation as and when they appear.[20] A more active role is appropriate at the feasibility testing stage. However, when evaluating effectiveness, researchers will ideally not engage in continuous quality improvement activities because these may compromise the external validity of the evaluation. Agreeing systems for communicating information to stakeholders at the outset of the study may help to avoid perceptions of undue interference or that the evaluator withheld important information.

Resources and staffing

When planning a process evaluation, evaluators need to ensure that there is sufficient expertise and experience to decide on, and achieve, its aims. A process evaluation team will often require expertise in quantitative and qualitative research methods. Process evaluations will often need to draw on expertise from a range of relevant disciplines including, for example, public health, primary care, epidemiology, sociology, and psychology. Sufficient resources are required to allow collection and analysis of large quantities of diverse data, bearing in mind that analysis of qualitative data is especially time consuming.

Relationships within evaluation teams

Process evaluation will typically form part of a study that includes evaluation of outcomes and possibly cost effectiveness. Some evaluators choose to separate process and outcome teams, while in other cases they are combined. Box 2 gives some pros and cons of each model. If the teams are separate effective communications are necessary to prevent duplication or conflict; with combined teams, there is a need for transparency about how this might influence the conduct and interpretation of the evaluation. Effective integration of evaluation components is more likely when members of a team respect and value each other's work, and when the overall study is overseen by a principal investigator who values integration.[21]

Designing and conducting a process evaluation

Describing the intervention and clarifying causal assumptions

A clear description of the intended intervention, how it will be implemented, and how it is expected to work, will ideally have been developed before evaluation. In such cases, designing a process evaluation will begin by reviewing these descriptions to decide what requires investigation. Any ambiguity over what the intervention is, or how it is intended to work, should be resolved with the intervention developers before the design of the process evaluation is finalised. Evaluators of NERS had limited involvement in the development of the intervention, which was a Welsh government policy initiative. Hence, when evaluation began, some ambiguity remained over the content of the intervention and how it was intended to work. Evaluators worked with intervention developers to resolve this ambiguity, but as this took place after the evaluation had started, the time available to develop robust measures of some key activities was limited.[8]

It is useful if interventions and their evaluations draw explicitly on existing theories so that these can be tested and refined. However, when an intervention's development is driven by other factors, such as experience or common sense, it is important to be open about

BOX 1: KEY RECOMMENDATIONS FOR PROCESS EVALUATION

Planning

- Carefully define the parameters of relationships with intervention developers or implementers
  - Balance the need for sufficiently good working relationships to allow close observation, against the need to remain credible as independent evaluators
  - Agree whether evaluators will take an active role in communicating findings as they emerge (and helping correct implementation challenges) or have a more passive role
- Ensure that the research team has the correct expertise. This may require:
  - Expertise in qualitative and quantitative research methods
  - Appropriate interdisciplinary theoretical expertise
- Decide the degree of separation or integration between process and outcome evaluation teams
  - Ensure effective oversight by a principal investigator who values all evaluation components
  - Develop good communication systems to minimise duplication and conflict between process and outcomes evaluations
  - Ensure that plans for integration of process and outcome data are agreed from the outset

Design and conduct

- Clearly describe the intervention and clarify causal assumptions (in relation to how it will be implemented, and the mechanisms through which it will produce change, in a specific context)
- Identify key uncertainties and systematically select the most important questions to address
  - Identify potential questions by considering the assumptions represented by the intervention
  - Agree scientific and policy priority questions by considering the evidence for intervention assumptions and consulting the evaluation team and policy or practice stakeholders
  - Identify previous process evaluations of similar interventions and consider whether it is appropriate to replicate aspects of them and build on their findings
- Select a combination of methods appropriate to the research questions:
  - Use quantitative methods to measure key process variables and allow testing of pre-hypothesised mechanisms of impact and contextual moderators
  - Use qualitative methods to capture emerging changes in implementation, experiences of the intervention and unanticipated or complex causal pathways, and to generate new theory
  - Balance collection of data on key process variables from all sites or participants with detailed data from smaller, purposively selected samples
  - Consider data collection at multiple time points to capture changes to the intervention over time

Analysis

- Provide descriptive quantitative information on fidelity, dose, and reach
- Consider more detailed modelling of variations between participants or sites in terms of factors such as fidelity or reach (eg, are there socioeconomic biases in who received the intervention?)
- Integrate quantitative process data into outcomes datasets to examine whether effects differ by implementation or prespecified contextual moderators, and test hypothesised mediators
- Collect and analyse qualitative data iteratively so that themes that emerge in early interviews can be explored in later ones
- Ensure that quantitative and qualitative analyses build upon one another (eg, qualitative data used to explain quantitative findings or quantitative data used to test hypotheses generated by qualitative data)
- Where possible, initially analyse and report process data before trial outcomes are known to avoid biased interpretation
- Transparently report whether process data are being used to generate hypotheses (analysis blind to trial outcomes), or for post-hoc explanation (analysis after trial outcomes are known)

Reporting

- Identify existing reporting guidance specific to the methods adopted
- Report the logic model or intervention theory and clarify how it was used to guide selection of research questions and methods
- Disseminate findings to policy and practice stakeholders
- If multiple journal articles are published from the same process evaluation ensure that each article makes clear its context within the evaluation as a whole:
  - Publish a full report comprising all evaluation components or a protocol paper describing the whole evaluation, to which reference should be made in all articles
  - Emphasise contributions to intervention theory or methods development to enhance interest to a readership beyond the specific intervention in question

this and clear about what these assumptions are, rather than trying to force an established theoretical framework to fit the intervention. Evaluators should also avoid focusing narrowly on inappropriate theories from a single discipline. For example, psychological theory may be useful for interventions that work at the individual level but is less useful when intervening with organisations or at wider social levels.[22]

Depicting the intervention in a logic model can help clarify causal assumptions.[23] Figure 2 gives an

doi: 10.1136/bmj.h1258 | *BMJ* 2015;350:h1258 | the bmj

BOX 2: SEPARATION OR INTEGRATION OF PROCESS EVALUATION AND OUTCOME EVALUATION TEAMS?

Arguments for separation

- Separation may reduce potential biases in analysis of outcomes data arising from feedback on the perceived functioning of the intervention
- In controlled trials, process evaluators cannot be blinded to treatment condition. Those collecting or analysing outcomes data ought to be blinded where possible
- Analysing process data without knowledge of trial outcomes prevents fishing for explanations and biasing interpretations. Although it may not always be practical to delay outcomes analysis until process analyses are complete, if separate researchers are responsible for each part it may be possible to conducted the analyses concurrently without biasing the results
- Process evaluation may produce data that would be hard for those with vested interests in the trial to analyse and report dispassionately
- If implementers or participants have concerns about a trial, a degree of separation from the trial may make it easier for process evaluators to build rapport and understand their concerns

Arguments for integration

- Process evaluators and outcomes evaluators will want to work together to ensure that data on implementation can be integrated into analysis of outcomes, or that data on emerging process issues can be integrated into trial data collections
- Data on intermediate outcomes and causal processes identified by process evaluators may inform integration of new measures into outcomes data collections
- If some relevant process measures are already being collected as part of the outcomes evaluation, it is important to avoid duplication of efforts and reduce measurement burden for participants
- One component of data collection should not compromise another. For example, if collection of process data is causing a high measurement burden for participants, this may lead to lower response to outcomes assessments

example for INCLUSIVE, a school based intervention that aimed to reduce bullying and improve student health by implementing "restorative practices" across the whole school.[24] The logic model was based on Markham and Aveyard's theory of human functioning and school organisation, which suggests that health benefits would be mediated by whether students were connected to their school's learning and community.[25] This led the authors to identify measures of commitment and belonging as intermediate outcomes.[26]

Learning from previous process evaluations

When designing a process evaluation, it is important to be mindful that the results may later be included in systematic reviews. Process evaluation will provide the information on implementation and context that Waters and colleagues argue is essential if reviews are to assist decision makers.[27] It is therefore helpful if process evaluations of similar interventions build on one another's findings, using comparable methods if possible, so that reviewers can make meaningful comparisons across studies.

Deciding core research questions

Process evaluations cannot expect to provide answers to all of the uncertainties of a complex intervention.[28] It is generally better to answer the most important questions well than to try to answer too many questions and do so unsatisfactorily. To identify core questions, evaluators may start by listing causal assumptions within the intervention manual or logic model and establishing which have the most limited evidence base. This can be done by reviewing the literature, consultation with policy and practice stakeholders, and discussions within the research team. Complex interventions are inherently unpredictable. Evaluators may therefore identify additional questions during the course of their evaluation. Hence, although clear focus from the outset is vital, process evaluations must be designed with sufficient flexibility and resources to allow important emerging questions to be addressed.

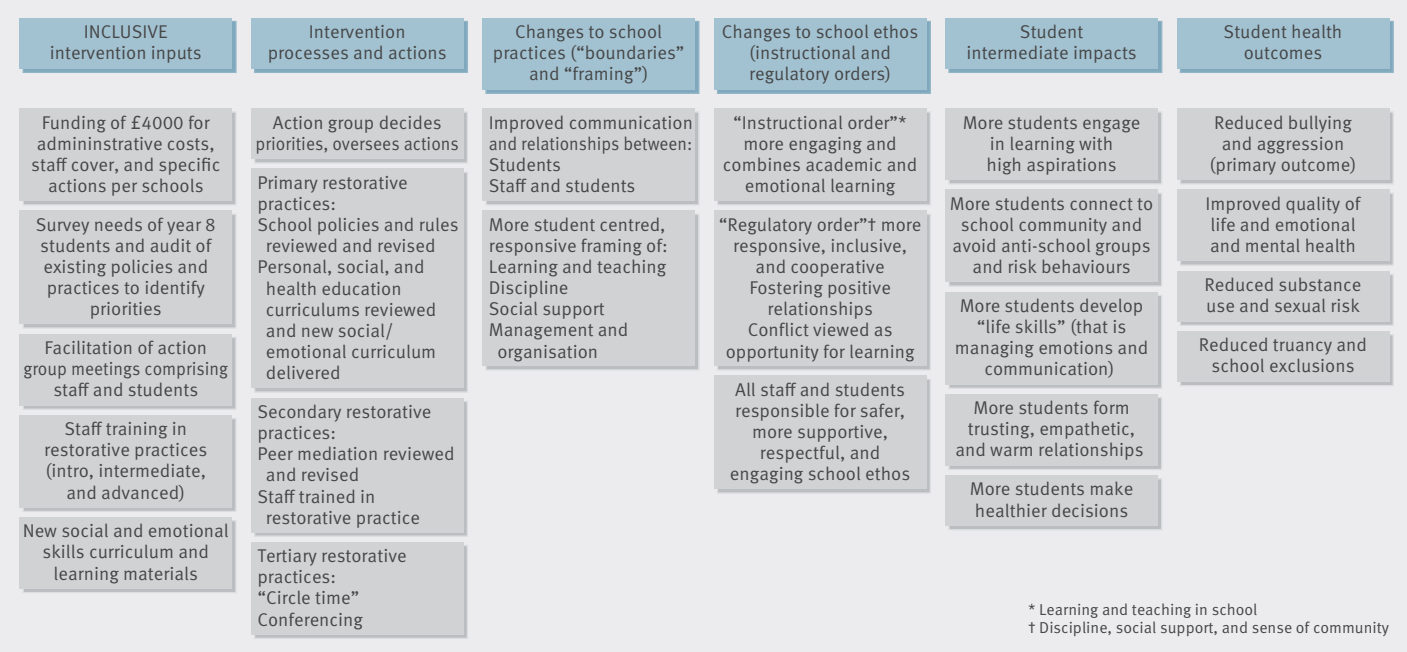


Fig 2 | **Logic model for the INCLUSIVE intervention to reduce violence and aggression in schools[24]**

Selecting methods

Figure 3 lists some common data collection and analysis methods adopted by process evaluations, the merits of which should be considered carefully in relation to the research questions. Process evaluation of complex interventions usually requires a combination of quantitative and qualitative methods, but their relative importance may vary according to the status of the evidence base or stage of the evaluation process. At the feasibility and piloting stage, basic quantitative measures of implementation may be combined with in-depth qualitative data to provide detailed understandings of intervention functioning on a small scale.

When evaluating effectiveness, collection of quantitative process measures to allow testing of hypothesised pathways or to measure contextual factors may be a priority. If directly relevant qualitative data are already available (for example, from an earlier feasibility study), evaluators may choose not to collect extensive qualitative process data while evaluating effectiveness. However, collecting additional qualitative data may still help in understanding issues arising from the movement from a small scale feasibility study to a larger scale evaluation involving greater diversity in implementers, settings, and participants.

Key methodological considerations include sampling and timing of data collection. Interviewing every implementer may not provide greater insights than interviewing a small well selected sample, and may lead to overwhelming volumes of data. Conducting observations in every site may be prohibitively expensive and unduly influence implementation. Conversely, there are dangers in collecting data from only a few sites in order to draw conclusions regarding the intervention as a whole.[28] Hence, when feasible, it is often useful to combine quantitative data on key process variables from all sites or participants with in-depth qualitative data from samples purposively selected along dimensions expected to influence the functioning of the intervention. Collecting data at multiple time points may be useful because interventions can suffer from teething problems which are rectified as the evaluation progresses.

Within the NERS process evaluation, quantitative measures included structured observations of audio recorded patient consultations. These were used to examine aspects of fidelity (such as consistency with motivational interviewing principles), and dose (such as the duration of consultations). Sociodemographic patterning in entry to the scheme (reach) was evaluated using routinely collected monitoring data.[8] Quantitative measures of hypothesised psychological mechanisms, including motivation for exercise and confidence, were collected as part of the trial.[18] Qualitative interviews were conducted with patients, exercise professionals, scheme coordinators, and health professionals. These focused on challenges in implementation across contexts and how NERS was perceived to work in practice.[8]

Analysis of process data, and integration of process and outcome data

Analysis of quantitative process data will usually begin with descriptive statistics relating to questions such as fidelity, dose, and reach. Subsequently, integrating quantitative process measures into outcomes datasets can help to understand how, for example, implementation variability affected outcomes (on-treatment analyses) and test hypotheses arising from qualitative analyses. Some argue that initial analysis of process data should be conducted before the outcomes analysis to avoid biased interpretation of process data.[29] If this model is followed, process data may provide prospective insights into why evaluators might subsequently expect to see positive or negative overall effects and generate hypotheses about how variability in outcomes may emerge.[30]

In the NERS process evaluation, implementation measures indicated that the intervention comprised a common core of health professional referrals to discounted, supervised, group based exercise. However, some activities, such as motivational interviewing and goal setting, were poorly delivered.[8] Nevertheless, qualitative data (analysed before trial outcomes were available) indicated that patient motivation was supported by other mechanisms, such as social support from other patients.[8] Subsequently, integration of quantitative measures of psychological change mechanisms with trial outcomes data indicated that significant improvement in physical activity was explained by change in

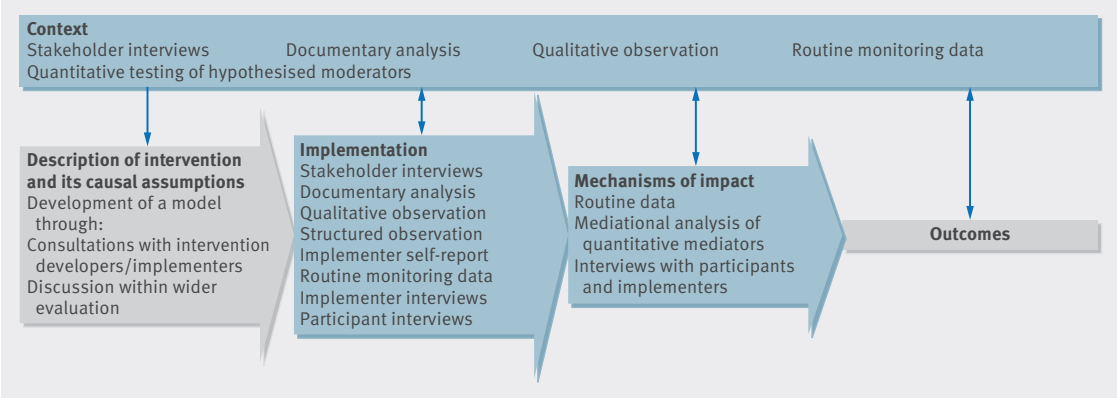


Fig 3 | Commonly used data collection and analysis methods for process evaluation

doi: 10.1136/bmj.h1258 | *BMJ* 2015;350:h1258 | the bmj

motivation for exercise.[18] Hence, the integration of qualitative and quantitative process data with trial outcomes helped to clarify complex causal pathways.

Reporting findings

Reporting guidelines for health research are available on the EQUATOR network website (www.equator-network.org/home), but such guidelines for process evaluations are challenging because they vary so much. Key considerations include reporting relations between quantitative and qualitative components, and the relation of the process evaluation to other evaluation components, such as outcomes or economic evaluation. It is also useful to report assumptions about how the intervention works (ideally in a logic model), and how these informed the selection of research questions and methods.[31] Reporting in the peer reviewed literature will often require multiple articles. To maintain sight of the broader picture, all journal articles should refer to other articles published from the study or to a protocol paper or report that clarifies how the component publications relate to the overall evaluation. When process evaluation has been conducted to interpret trial outcomes, interpretation needs to be clear in the published papers, with process evaluation data linked, in discussion, to trial outcomes. It is also important to report in lay formats for people who delivered the intervention or who will be making decisions about its future implementation.

Contributors: GM led the development of the guidance, wrote the first draft of the article, and the full guidance document which it describes, and integrated contributions from the author group into subsequent drafts. JB was the lead applicant for the funding to conduct the work and chaired the author group. All authors contributed to the design and content of the guidance and subsequent drafts of the paper. GM acts as guarantor.

Funding: The work was funded by the MRC Population Health Science Research Network (PHSRN45).

Competing interests: All authors have read and understood BMJ policy on declaration of interests and have no relevant interests to declare.

Provenance and peer review: Not commissioned; externally peer reviewed.

1 Craig P, Dieppe P, Macintyre S, et al. Developing and evaluating complex interventions: the new Medical Research Council guidance. *BMJ* 2008;337:a1655.
2 Campbell M, Fitzpatrick R, Haines A, et al. Framework for design and evaluation of complex interventions to improve health. *BMJ* 2000;321:694–96.
3 Craig P, Dieppe P, Macintyre S, et al. Developing and evaluating complex interventions: new guidance: MRC, 2008.
4 Moore G, Audrey S, Barker M, et al. Process evaluation in complex public health intervention studies: the need for guidance. *J Epidemiol Community Health* 2014;68:101–02.
5 Fretheim A, Flottorp S, Oxman AD. It is a capital mistake to theorize before one has data: a response to Eccle's criticism of the OFF theory of research utilization. *J Clin Epidemiol* 2005;58:119–20.
6 De Silva M, Breuer E, Lee L, et al. Theory of change: a theory-driven approach to enhance the Medical Research Council's framework for complex interventions. *Trials* 2014;15:267.
7 Steckler A, Linnan L, editors. Process evaluation for public health interventions and research. Jossey-Bass, 2002.
8 Moore GF, Raisanen L, Moore L, et al. Mixed-method process evaluation of the Welsh National Exercise Referral Scheme. *Health Education* 2013;113:476–501.
9 Hawe P, Shiell A, Riley T. Complex interventions: how "out of control" can a randomised controlled trial be? *BMJ* 2004;328:1561–63.
10 Bumbarger B, Perkins D. After randomised trials: issues related to dissemination of evidence-based interventions. *J Children Serv* 2008;3:55–64.
11 Carroll C, Patterson M, Wood S, et al. A conceptual framework for implementation fidelity. *Implement Sci* 2007;2:40.
12 Montgomery P, Underhill K, Gardner F, et al. The Oxford Implementation Index: a new tool for incorporating implementation data into systematic reviews and meta-analyses. *J Clin Epidemiol* 2013;66:874–82.
13 Glasgow RE, Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. *Am J Public Health* 1999;89:1322–27.
14 Grant A, Treweek S, Dreischulte T, et al. Process evaluations for cluster-randomised trials of complex interventions: a proposed framework for design and reporting. *Trials* 2013;14:15.
15 Bonell C, Fletcher A, Morton M, et al. Realist randomised controlled trials: a new approach to evaluating complex public health interventions. *Soc Sci Med* 2012;75:2299–306.
16 Shiell A, Hawe P, Gold L. Complex interventions or complex systems? Implications for health economic evaluation. *BMJ* 2008;336:1281–83.
17 Pawson R, Tilley N. *Realistic evaluation*. Sage, 1997.
18 Littlecott H, Moore G, Moore L, et al. Psychosocial mediators of change in physical activity in the Welsh national exercise referral scheme: secondary analysis of a randomised controlled trial. *Int J Behav Nutr Physical Activity* 2014;11:109.
19 Moore GF, Moore L, Murphy S. Integration of motivational interviewing into practice in the national exercise referral scheme in Wales: a mixed methods study. *Behav Cog Psychother* 2012;40:313–30.
20 Audrey S, Holliday J, Parry-Langdon N, et al. Meeting the challenges of implementing process evaluation within randomized controlled trials: the example of ASSIST (A Stop Smoking in Schools Trial). *Health Educ Res* 2006;21:366–77.
21 O'Cathain A, Murphy E, Nicholl J. Multidisciplinary, interdisciplinary, or dysfunctional? Team working in mixed-methods research. *Qual Health Res* 2008;18:1574–85.
22 Hawe P, Shiell A, Riley T. Theorising Interventions as Events in Systems. *Am J Community Psychol* 2009;43:267–76.
23 Kellogg Foundation WK. Logic model development guide. W K Kellogg Foundation, 2004.
24 Bonell C, Fletcher, A, Fitzgerald-Yau, N, et al. Initiating change locally in bullying and aggression through the school environment (INCLUSIVE): pilot randomised controlled trial. *Health Technol Assess* (forthcoming).
25 Markham WA, Aveyard P. A new theory of health promoting schools based on human functioning, school organisation and pedagogic practice. *Soc Sci Med* 2003;56:1209–20.
26 Sawyer MG, Pfeiffer S, Spence SH, et al. School based prevention of depression: a randomised controlled study of the beyondblue schools research initiative. *J Child Psychol Psychiatry* 2010;51:199–209.
27 Waters E, Hall BJ, Armstrong R, et al. Essential components of public health evidence reviews: capturing intervention complexity, implementation, economics and equity. *J Public Health* 2011;33:462–65.
28 Munro A, Bloor M. Process evaluation: the new miracle ingredient in public health research? *Qualitative Research* 2010;10:699–713.
29 Oakley A, Strange V, Bonell C, et al. Health services research—process evaluation in randomised controlled trials of complex interventions. *BMJ* 2006;332:413–6.
30 Mermelstein R. Moving tobacco prevention outside the classroom. *Lancet* 2008;371:1556–57.
31 Armstrong R, Waters E, Moore L, et al. Improving the reporting of public health intervention research: advancing TREND and CONSORT. *J Public Health* 2008;30:103–9.

Framework for design and evaluation of complex interventions to improve health

Michelle Campbell, Ray Fitzpatrick, Andrew Haines, Ann Louise Kinmonth, Peter Sandercock, David Spiegelhalter, Peter Tyrer

Office of the President, Medical Research Council of Canada, 1600 Scott Street, Ottawa, Ontario, Canada K1 OW9

Michelle Campbell *senior policy analyst*

Division of Public Health and Primary Health Care, Institute of Health Sciences, University of Oxford, Oxford OX3 7LF

Ray Fitzpatrick *professor of public health and primary care*

Department of Primary Care and Population Sciences, Royal Free and University College Medical School, London NW3 2PF

Andrew Haines *professor of primary health care*

General Practice and Primary Care Research Unit, Department of Public Health and Primary Care, Institute of Public Health, Cambridge CB2 2SR

Ann Louise Kinmonth *professor of general practice*

Neuroscience Trials Unit, Department of Clinical Neurosciences, Western General Hospitals NHS Trust, Edinburgh EH4 2XU

Peter Sandercock *professor of medical neurology*

continued over

BMJ 2000;321:694–6

Randomised controlled trials are widely accepted as the most reliable method of determining effectiveness, but most trials have evaluated the effects of a single intervention such as a drug. Recognition is increasing that other, non-pharmacological interventions should also be rigorously evaluated.[1–3] This paper examines the design and execution of research required to address the additional problems resulting from evaluation of complex interventions—that is, those "made up of various interconnecting parts."[4] The issues dealt with are discussed in a longer Medical Research Council paper (www.mrc.ac.uk/complex_packages.html). We focus on randomised trials but believe that this approach could be adapted to other designs when they are more appropriate.

Summary points

- Complex interventions are those that include several components
- The evaluation of complex interventions is difficult because of problems of developing, identifying, documenting, and reproducing the intervention
- A phased approach to the development and evaluation of complex interventions is proposed to help researchers define clearly where they are in the research process
- Evaluation of complex interventions requires use of qualitative and quantitative evidence

Challenges of trials of complex interventions

There are specific difficulties in defining, developing, documenting, and reproducing complex interventions that are subject to more variation than a drug. A typical example would be the design of a trial to evaluate the benefits of specialist stroke units. Such a trial would have to consider the expertise of various health professionals as well as investigations, drugs, treatment guidelines, and arrangements for discharge and follow up. Stroke units may also vary in terms of organisation, management, and skill mix. The active components of the stroke unit may be difficult to specify, making it difficult to replicate the intervention. The box gives other examples of complex interventions.

Examples of complex interventions

Service delivery and organisation:
- Stroke units
- Hospital at home

Interventions directed at health professionals' behaviour:
- Strategies for implementing guidelines
- Computerised decision support

Community interventions:
- Community based programmes to prevent heart disease
- Community development approaches to improve health

Group interventions:
- Group psychotherapies or behavioural change strategies
- School based interventions—for example, to reduce smoking or teenage pregnancy

Interventions directed at individual patients:
- Cognitive behavioural therapy for depression
- Health promotion interventions to reduce alcohol consumption or support dietary change

Framework for trials of complex interventions

Problems often arise in the evaluation of complex interventions because researchers have not fully defined and developed the intervention. It is useful to consider the process of development and evaluation of such interventions as having several distinct phases. These can be compared with the sequential phases of drug development (fig 1) or may be seen as more iterative (fig 2). Either way a phased approach separates the different questions being asked.

Progression from one phase to another may not be linear. In many cases an iterative process occurs—for example, if an exploratory trial finds that a complex intervention is unacceptable to potential recipients, the theoretical basis and components of the intervention may have to be re-examined. Preliminary work is often essential to establish the probable active

MRC Biostatistics Unit, Institute of Public Health, Cambridge CB2 2SR
David Spiegelhalter *senior scientist*
Department of Public Mental Health, Imperial College of Science, Technology, and Medicine, St Mary's Campus, London W2 1PD
Peter Tyrer *professor of community psychiatry*
Correspondence to: R Fitzpatrick raymond.fitzpatrick@nuffield.ox.ac.uk

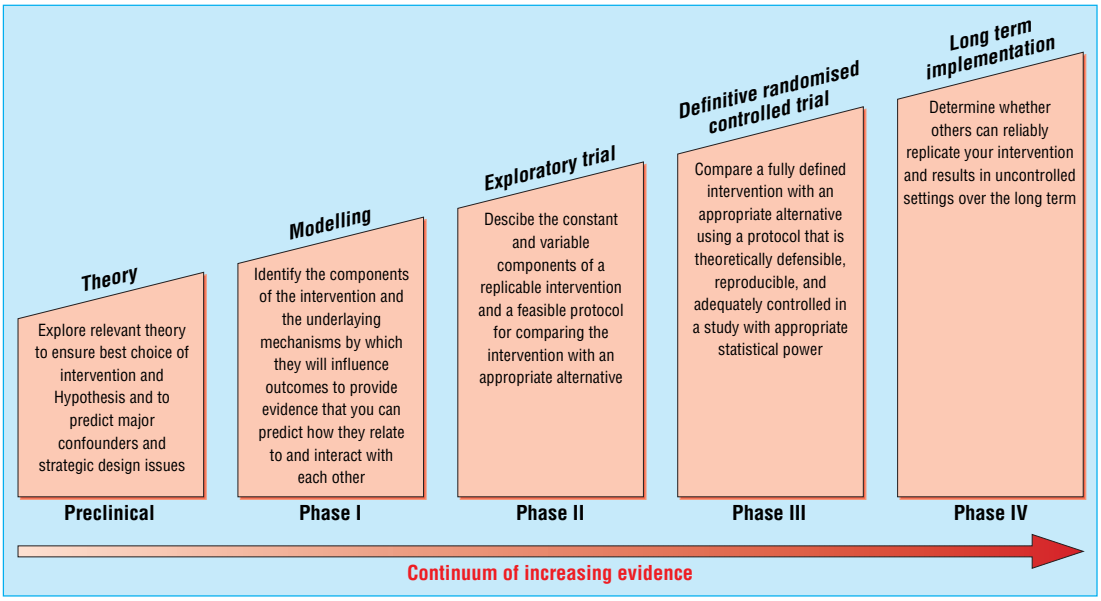


Fig 1 Sequential phases of developing randomised controlled trials of complex interventions

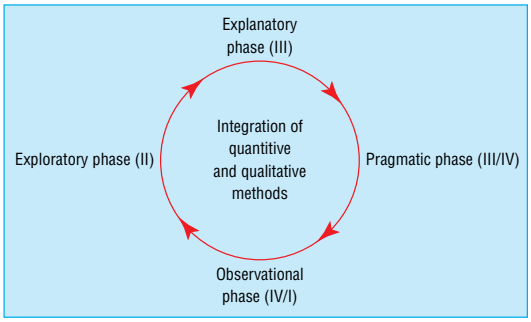


Fig 2 Iterative view of development of randomised controlled trials of complex interventions

components of the intervention so that they can be delivered effectively during the trial. Identifying which stage of development has been reached in specifying the intervention and outcome measures will give researchers and funding bodies reasonable confidence that an appropriately designed and relevant study is being proposed.

Preclinical or theoretical phase

The first step is to identify the evidence that the intervention might have the desired effect. This may come from disciplines outside the health sciences (such as theories of organisational change). Review of the theoretical basis for an intervention may lead to changes in the hypothesis and improved specification of potentially active ingredients. In addition, previous studies may have provided some empirical evidence—for example, an intervention may have been found effective for a closely related condition or in another country with a different organisation of health care.[5]

Phase I: defining components of the intervention

Modelling or simulation techniques can improve understanding of the components of an intervention and their interrelationships. Qualitative testing through focus groups, preliminary surveys, or case studies can also help define relevant components. Descriptive studies may help to delineate variants of a service. For example, hospital at home schemes vary in purpose. Some are designed to hasten hospital discharge, others to avoid hospital admissions, and yet others to provide palliative care in the home.[6]

Qualitative research can also be used to show how the intervention works and to find potential barriers to change in trials that seek to alter patient or professional behaviour.[7] For example, if health professionals see the main barrier to changing their practice as being lack of time or resources, an intervention that focuses only on improving their knowledge will not work.

Phase II: defining trial and intervention design

Acceptability and feasibility

In phase II the information gathered in phase I is used to develop the optimum intervention and study design. This often involves testing the feasibility of delivering the intervention and acceptability to providers and patients. Different versions of the intervention may need to be tested or the intervention may have to be adapted to achieve optimal effectiveness—for example, if the proposed intensity and duration of the intervention are found to be unacceptable to participants.

It is also important to test for evidence of a learning curve, leading to improved performance of the intervention over time. If a learning curve exists a run-in period might be needed before formal recruitment to the trial to ensure that the intervention is provided effectively.

The exploratory trial is also an opportunity to determine the consistency with which the intervention is delivered. Consultations could be audio or video taped to give feedback of performance to providers together with training to promote consistency.

Defining the control intervention

The content of the comparative arm (control group) of the main trial will be decided during the preparatory phase. It may be an alternative package of care, standard care, or placebo. Although standard practice is often an appropriate control, it can be as complex as the intervention being evaluated and may change with time. It is thus important to monitor the care that is being delivered to the control group. The use of a no treatment control group may be unacceptable to patients. One possible solution is a randomised waiting list study in which all participants ultimately receive the intervention.

Designing the main trial

The exploratory phase should ideally be randomised to allow assessment of the size of the effect. This initial assessment will provide a sound basis for calculating sample sizes for the main trial. Other design variables can also be established in an exploratory trial.

Outcomes

Outcome measures for the main trial will also generally be piloted during the exploratory phase. Investigators should include outcomes that not only are relevant to patients with the disease or condition being studied but also encompass measures of wider relevance to the health system, including economic measures.[8] Collection of data to assess a full range of costs to patients, carers, and society adds considerably to the workload and costs of researchers and may challenge the feasibility of a trial. Strategic choice of outcomes is therefore needed.[9]

An important decision in trials of complex interventions is whether health outcome needs to be assessed. For studies such as those evaluating strategies to change professional behaviour, it may be sufficient to show that the intervention changed behaviour, provided that clear evidence exists that the changed behaviour—for example, prescribing particular treatments—is effective.

Phase III: methodological issues for main trial

The main trial will need to address the issues normally posed by randomised controlled trials, such as sample size, inclusion and exclusion criteria, and methods of randomisation, as well as the challenges of complex interventions. Individual randomisation may not always be feasible or appropriate. For example, cluster randomisation is often used for trials of interventions directed at a practice or hospital team.[10 11] Randomised incomplete block designs have also been used to evaluate different approaches to promoting change in professional behaviour.[12]

It is often not possible to conceal allocation of treatment from the patient, practitioner, and researcher in complex intervention trials. The potential biases of unblinded trials therefore have to be taken into account. Dissimilar levels of patient commitment between intervention and control groups may cause differential dropout, making interpretation of results difficult. When patients have strong preferences, a preference trial design may be used; patients without strong preferences are randomised as usual but those with strong preferences receive their preferred treatment.[13] The results of such trials can, however, be difficult to interpret.

The findings of trials of complex interventions are more generalisable if they are performed in the setting in which they are most likely to be implemented. Eligibility criteria must not lead to the exclusion of patients—for example, on the grounds of age—who constitute a substantial portion of those to whom the intervention is likely to be offered when implemented in the health system. Poor recruitment to a trial can also raise doubts about generalisability.

Qualitative study of the processes of implementation of interventions in study arms of the main trial can further show the validity of findings.[14]

Phase IV: promoting effective implementation

The purpose of the final phase is to examine the implementation of the intervention into practice, paying particular attention to the rate of uptake, the stability of the intervention, any broadening of subject groups, and the possible existence of adverse effects. As in the case of drug trials, this might be carried out by long term surveillance, although currently there is no established mechanism for funding such activities.

Conclusions

Trials of complex interventions are of increasing importance because of the drive to provide the most cost effective health care. Although these trials pose substantial challenges to investigators, the use of an iterative phased approach that harnesses qualitative and quantitative methods should lead to improved study design, execution, and generalisability of results.

We thank the participants at a MRC workshop on complex interventions for their contribution. This article represents the views of the authors and does not represent the official view of the Medical Research Council.

Competing interests: None declared.

1 Friedli K, King MB. Psychological treatments and their evaluation. *Int Rev Psychiatry* 1998;10:123-6.
2 Stephenson J, Imrie J. Why do we need randomised controlled trials to assess behavioural interventions? *BMJ* 1998;316:611-3.
3 Buchwald H. Surgical procedures and devices should be evaluated in the same way as medical therapy. *Controlled Clinical Trials* 1997;18:478-87.
4 *Collins English dictionary*. London: Collins, 1979.
5 Haines A, Iliffe S. Innovations in services and the appliance of science. *BMJ* 1995;310:815-6.
6 Shepperd S, Iliffe S. The effectiveness of hospital at home compared with in-patient hospital care: a systematic review. *J Pub Health Med* 1998; 20:344-50.
7 Haynes RB, Haines A. Barriers and bridges to evidence-based practice. *BMJ* 1998;317:273-6.
8 Fitzpatrick R, Davey C, Buxton MJ, Jones DR. Evaluating patient based outcome measures for use in clinical trials. *Health Technology Assessment* 1998;2(14):1-74.
9 Mannheim L. Health services research clinical trials: issues in the evaluation of economic cost and benefits. *Controlled Clinical Trials* 1999;19: 149-58.
10 Edwards S, Braunholtz D, Stevens A, Lilford R, Ethical issues in the design and conduct of cluster RCTs. *BMJ* 1999;318:1407-9.
11 Ukoumunne OC, Gulliford MC, Chinn S, Sterne JAC, Burney PGJ. Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review. *Health Technology Assessment* 1999;3(5):1-110.
12 Grimshaw J, Freemantle N, Wallace S, Russell I, Hurwitz B, Watt I, et al. Developing and implementing clinical practice guidelines. *Qual Health Care* 1995;4:55-64.
13 Brewin CR, Bradley C. Patient preferences and randomised controlled trials. *BMJ* 1989;289:313-5.
14 Bradley F, Wiles R, Kinmonth A, Mant D, Gantley M. Development and evaluation of complex interventions in health services research: case study of the Southampton heart integrated care project (SHIP). *BMJ* 1999;318:711-5.

(Accepted 31 May 2000)

BMJ 2008;337:a1655 doi: 10.1136/bmj.a1655 (Published 29 September 2008)

RESEARCH METHODS & REPORTING

Developing and evaluating complex interventions: the new Medical Research Council guidance

Evaluating complex interventions is complicated. The Medical Research Council's evaluation framework (2000) brought welcome clarity to the task. Now the council has updated its guidance

Peter Craig *programme manager*[1], Paul Dieppe *professor*[2], Sally Macintyre *director*[3], Susan Michie *professor*[4], Irwin Nazareth *director*[5], Mark Petticrew *professor*[6]

[1]MRC Population Health Sciences Research Network, Glasgow G12 8RZ; [2]Nuffield Department of Orthopaedic Surgery, University of Oxford, Nuffield Orthopaedic Centre, Oxford OX3 7LD; [3]MRC Social and Public Health Sciences Unit, Glasgow G12 8RZ; [4]Centre for Outcomes Research and Effectiveness, University College London, London WC1E 7HB; [5]MRC General Practice Research Framework, London NW1 2ND; [6]Public and Environmental Health Research Unit, Department of Public Health and Policy, London School of Hygiene and Tropical Medicine, London WC1E 7HT

Complex interventions are widely used in the health service, in public health practice, and in areas of social policy that have important health consequences, such as education, transport, and housing. They present various problems for evaluators, in addition to the practical and methodological difficulties that any successful evaluation must overcome. In 2000, the Medical Research Council (MRC) published a framework[1] to help researchers and research funders to recognise and adopt appropriate methods. The framework has been highly influential, and the accompanying *BMJ* paper is widely cited.[2] However, much valuable experience has since accumulated of both conventional and more innovative methods. This has now been incorporated in comprehensively revised and updated guidance recently released by the MRC (www.mrc.ac.uk/complexinterventionsguidance). In this article we summarise the issues that prompted the revision and the key messages of the new guidance.

Revisiting the 2000 MRC framework

As experience of evaluating complex interventions has accumulated since the 2000 framework was published, interest in the methodology has also grown. Several recent papers have identified limitations in the framework, recommending, for example, greater attention to early phase piloting and development work,[3] a less linear model of evaluation process,[4] integration of process and outcome evaluation,[5] recognition that complex interventions may work best if they are tailored to local contexts rather than completely standardised,[6] and greater use of the insights provided by the theory of complex adaptive systems.[7]

A workshop held by the MRC Population Health Sciences Research Network to consider whether and how the framework should be updated likewise recommended the inclusion of a model of the evaluation process less closely to tied to the phases of drug development; more guidance on how to approach the development, reporting, and implementation of complex interventions; and greater attention to the contexts in which interventions take place. It further recommended consideration of alternatives to randomised trials, and of highly complex or non-health sector interventions to which biomedical methods may not be applicable, and more evidence and examples to back up and illustrate the recommendations. The new guidance addresses these issues in depth, and here we set out the key messages.

What are complex interventions?

Complex interventions are usually described as interventions that contain several interacting components, but they have other characteristics that evaluators should take into account (box 1). There is no sharp boundary between simple and complex interventions. Few interventions are truly simple, but the number of components and range of effects may vary widely. Some highly complex interventions, such as the Sure Start intervention to support families with young children in deprived communities,[8] may comprise a set of individually complex interventions.

How these characteristics are dealt with will depend on the aims of the evaluation. A key question in evaluating complex interventions is whether they are effective in everyday practice (box 2).[9] It is therefore important to understand the whole range of effects and how they vary, for example, among recipients or between sites. A second key question in evaluating complex interventions is how the intervention works: what are the active ingredients and how are they exerting their effect? Answers to this kind of question are needed to design more effective

Correspondence to: P Craig peter@sphsu.mrc.ac.uk

Summary points

- The Medical Research Council guidance for the evaluation of complex interventions has been revised and updated
- The process of developing and evaluating a complex intervention has several phases, although they may not follow a linear sequence
- Experimental designs are preferred to observational designs in most circumstances, but are not always practicable
- Understanding processes is important but does not replace evaluation of outcomes
- Complex interventions may work best if tailored to local circumstances rather than being completely standardised
- Reports of studies should include a detailed description of the intervention to enable replication, evidence synthesis, and wider implementation

Box 1 What makes an intervention complex?

- Number of interacting components within the experimental and control interventions
- Number and difficulty of behaviours required by those delivering or receiving the intervention
- Number of groups or organisational levels targeted by the intervention
- Number and variability of outcomes
- Degree of flexibility or tailoring of the intervention permitted

interventions and apply them appropriately across group and setting.[10]

Development, evaluation, and implementation

The 2000 framework characterised the process of development through to implementation of a complex intervention in terms of the phases of drug development. Although it is useful to think in terms of phases, in practice these may not follow a linear or even a cyclical sequence (figure⇓).[4]

Best practice is to develop interventions systematically, using the best available evidence and appropriate theory, then to test them using a carefully phased approach, starting with a series of pilot studies targeted at each of the key uncertainties in the design, and moving on to an exploratory and then a definitive evaluation. The results should be disseminated as widely and persuasively as possible, with further research to assist and monitor the process of implementation.

In practice, evaluation takes place in a wide range of settings that constrain researchers' choice of interventions to evaluate and their choice of evaluation methods. Ideas for complex interventions emerge from various sources, which may greatly affect how much leeway the researcher has to modify the intervention, to influence the way it is implemented, or to adopt an ideal evaluation design.[8] Evaluation may take place alongside large scale implementation, rather than starting beforehand. Strong evidence may be ignored or weak evidence taken up, depending on its political acceptability or fit with other ideas about what works.[11]

Researchers need to consider carefully the trade-off between the importance of the intervention and the value of the evidence that can be gathered given these constraints. In an evaluation of the health impact of a social intervention, such as a programme of housing improvement, the researcher may have no say in what the intervention consists of and little influence over how or when the programme is implemented, limiting the scope to undertake development work or to determine allocation. Experimental methods are becoming more widely accepted as methods to evaluate policy,[12] but there may be political or ethical objections to using them to assess health effects, especially if the intervention provides important non-health benefits.[13] Given the cost of such interventions, evaluation should still be considered—the best available methods, even if they are not optimal in terms of internal validity, may yield useful results.[14]

If non-experimental methods are used, researchers should be aware of their limitations and interpret and present the findings with due caution. Wherever possible, evidence should be combined from different sources that do not share the same weaknesses.[15] Researchers should be prepared to explain to decision makers the need for adequate development work, the pros and cons of experimental and non-experimental approaches, and the trade-offs involved in settling for weaker methods. They should be prepared to challenge decision makers when interventions of uncertain effectiveness are being implemented in a way that would make strengthening the evidence through a rigorous evaluation difficult, or when a modification of the implementation strategy would open up the possibility of a much more informative evaluation.

Developing a complex intervention

Identifying existing evidence—Before a substantial evaluation is undertaken, the intervention must be developed to the point where it can reasonably be expected to have a worthwhile effect. The first step is to identify what is already known about similar interventions and the methods that have been used to evaluate them. If there is no recent, high quality systematic review of the relevant evidence, one should be conducted and updated as the evaluation proceeds.

Identifying and developing theory—The rationale for a complex intervention, the changes that are expected, and how change is to be achieved may not be clear at the outset. A key early task is to develop a theoretical understanding of the likely process of change by drawing on existing evidence and theory, supplemented if necessary by new primary research. This should be done whether the researcher is developing the intervention or evaluating one that has already been developed.

Modelling process and outcomes—Modelling a complex intervention before a full scale evaluation can provide important information about the design of both the intervention and the evaluation. A series of studies may be required to progressively refine the design before embarking on a full scale evaluation. Developers of a trial of physical activity to prevent type 2 diabetes adopted a causal modelling approach that included a range of primary and desk based studies to design the intervention, identify suitable measures, and predict long term outcomes.[3] Another useful approach is a prior economic evaluation.[16] This may identify weaknesses and lead to refinements, or even show that a full scale evaluation is

Box 2 Developing and evaluating complex studies

- A good theoretical understanding is needed of how the intervention causes change, so that weak links in the causal chain can be identified and strengthened
- Lack of effect may reflect implementation failure (or teething problems) rather than genuine ineffectiveness; a thorough process evaluation is needed to identify implementation problems
- Variability in individual level outcomes may reflect higher level processes; sample sizes may need to be larger to take account of the extra variability and cluster randomised designs considered
- A single primary outcome may not make best use of the data; a range of measures will be needed and unintended consequences picked up where possible
- Ensuring strict standardisation may be inappropriate; the intervention may work better if a specified degree of adaptation to local settings is allowed for in the protocol

unwarranted. A modelling exercise to prepare for a trial of falls prevention in elderly people showed that the proposed system of screening and referral was highly unlikely to be cost effective and informed the decision not to proceed with the trial.[17]

Assessing feasibility

Evaluations are often undermined by problems of acceptability, compliance, delivery of the intervention, recruitment and retention, and smaller than expected effect sizes that could have been predicted by thorough piloting.[18] A feasibility study for an evaluation of an adolescent sexual health intervention in rural Zimbabwe found that the planned classroom based programme was inappropriate, given cultural norms, teaching styles, and relationships between teachers and pupils in the country, and it was replaced by a community based programme.[19] As well as illustrating the value of feasibility testing, the example shows the importance of understanding the context in which interventions take place.

A pilot study need not be a scale model of the planned evaluation but should examine the key uncertainties that have been identified during development. Pilot studies for a trial of free home insulation suggested that attrition might be high, so the design was amended such that participants in the control group received the intervention after the study.[20] Pilot study results should be interpreted cautiously when making assumptions about the numbers required when the evaluation is scaled up. Effects may be smaller or more variable and response rates lower when the intervention is rolled out across a wider range of settings.

Evaluating a complex intervention

There are many study designs to choose from, and different designs suit different questions and circumstances. Researchers should beware of blanket statements about what designs are suitable for what kind of intervention and choose on the basis of specific characteristics of the study, such as expected effect size and likelihood of selection or allocation bias. Awareness of the whole range of experimental and non-experimental approaches should lead to more appropriate methodological choices.

Assessing effectiveness

Randomisation should always be considered because it is the most robust method of preventing selection bias. If a conventional parallel group randomised trial is not appropriate, other randomised designs should be considered (box 3).

If an experimental approach is not feasible, because the intervention is irreversible, necessarily applies to the whole population, or because large scale implementation is already under way, a quasi-experimental or an observational design may be considered. In some circumstances, randomisation may be unnecessary and other designs preferable,[21 22] but the conditions under which observational methods can yield reliable estimates of effect are limited (box 4).[23] Successful examples, such as the evaluation of legislation to restrict access to means of suicide,[24] reduce air pollution,[25] or ban smoking in public places,[26] tend to occur where interventions have rapid, large effects.

Measuring outcomes

Researchers need to decide which outcomes are most important, which are secondary, and how they will deal with multiple outcomes in the analysis. A single primary outcome and a small number of secondary outcomes are the most straightforward for statistical analysis but may not represent the best use of the data or provide an adequate assessment of the success or otherwise of an intervention that has effects across a range of domains. It is important also to consider which sources of variation in outcomes matter and to plan appropriate subgroup analyses.

Long term follow-up may be needed to determine whether outcomes predicted by interim or surrogate measures do occur or whether short term changes persist. Although uncommon, such studies can be highly informative. Evaluation of a preschool programme for disadvantaged children showed that, as well as improved educational attainment, there was a range of economic and social benefits at ages 27 and 40.[27]

Understanding processes

Process evaluations, which explore the way in which the intervention under study is implemented, can provide valuable insight into why an intervention fails or has unexpected consequences, or why a successful intervention works and how it can be optimised. A process evaluation nested inside a trial can be used to assess fidelity and quality of implementation, clarify causal mechanisms, and identify contextual factors associated with variation in outcomes.[5] However, it is not a substitute for evaluation of outcomes. A process evaluation[28] carried out in connection with a trial of educational visits to encourage general practitioners to follow prescribing guidelines[29] found that the visits were well received and recall of the guidelines was good, yet there was little change in prescribing behaviour, which was constrained by other factors such as patients' preferences and local hospital policy.

Fidelity is not straightforward in relation to complex interventions.[30] In some evaluations, such as those seeking to identify active ingredients within a complex intervention, strict standardisation may be required and controls put in place to limit variation in implementation.[31] But some interventions are designed to be adapted to local circumstances. In a trial of a school based intervention to promote health and wellbeing, schools were encouraged to use a standardised process to develop strategies which suited them rather than adopt a fixed curriculum, resulting in widely varied practice between schools.[32] The key is to be clear about how much change or adaptation is

Box 3 Experimental designs for evaluating complex interventions

Individually randomised trials—Individuals are randomly allocated to receive either an experimental intervention or an alternative such as standard treatment, a placebo, or remaining on a waiting list. Such trials are sometimes dismissed as inapplicable to complex interventions, but there are many variants, and often solutions can be found to the technical and ethical problems associated with randomisation

Cluster randomised trials are one solution to the problem of contamination of the control group, leading to biased estimates of effect size, in trials of population level interventions. Groups such as patients in a general practice or tenants in a housing scheme are randomly allocated to the experimental or control intervention

Stepped wedge designs may be used to overcome practical or ethical objections to experimentally evaluating an intervention for which there is some evidence of effectiveness or which cannot be made available to the whole population at once. It allows a trial to be conducted without delaying roll-out of the intervention. Eventually, the whole population receives the intervention, but with randomisation built into the phasing of implementation

Preference trials and randomised consent designs—Practical or ethical obstacles to randomisation can sometimes be overcome by using non-standard designs. When patients have strong preferences among treatments, basing treatment allocation on patients' preferences or randomising patients before seeking consent may be appropriate.

N of 1 designs—Conventional trials aim to estimate the average effect of an intervention in a population. N of 1 trials, in which individuals undergo interventions with the order or scheduling decided at random, can be used to assess between and within person change and to investigate theoretically predicted mediators of that change

Box 4 Choosing between randomised and non-randomised designs

Size and timing of effects—Randomisation may be unnecessary if the effects of the intervention are so large or immediate that confounding or underlying trends are unlikely to explain differences in outcomes before and after exposure. It may be inappropriate—for example, on grounds of cost or delay—if the changes are very small or take a long time to appear. In these circumstances a non-randomised design may be the only feasible option, in which case firm conclusions about the impact of the intervention may be unattainable

Likelihood of selection bias—Randomisation is needed if exposure to the intervention is likely to be associated with other factors that influence outcomes. Post-hoc adjustment is a second best solution; its effectiveness is limited by errors in the measurement of the confounding variables and the difficulty of dealing with unknown or unmeasured confounders

Feasibility and acceptability of experimentation—Randomisation may be impractical if the intervention is already in widespread use, or if key decisions about how it will be implemented have already been taken, as is often the case with policy changes and interventions whose effect on health is secondary to their main purpose

Cost—If an experimental study is feasible and would provide more reliable information than an observational study but would also cost more, the additional cost should be weighed against the value of having better information

permissible and to record variations in implementation so that fidelity can be assessed in relation the degree of standardisation required by the study protocol.

Variability in implementation, preplanned or otherwise, makes it important that both process and outcome evaluations are reported fully and that a clear description of the intervention is provided to enable replication and synthesis of evidence.[33] This has been a weakness of the reporting of complex intervention studies in the past,[34] but the availability of a comprehensive range of reporting guidelines, now covering non-drug trials[35] and observational studies[36] and accessible through a single website (www.equator-network.org) should lead to improvement.

Conclusions

We recognise that many issues surrounding evaluation of complex interventions are still debated, that methods will continue to develop, and that practical applications will be found for some of the newer theories. We do not intend the revised guidance to be prescriptive but to help researchers, funders, and other decision makers to make appropriate methodological and practical choices. We have primarily aimed our messages at researchers, but publishers, funders, and commissioners of research also have an important part to play. Journal editors should insist on high and consistent standards of reporting. Research funders should be prepared to support developmental studies before large scale evaluations. The key message for policy makers is the need to consider evaluation requirements in the planning of new initiatives, and wherever possible to allow for an experimental or a high quality non-experimental approach to the evaluation of initiatives when there is uncertainty about their effectiveness.

Contributors: PD had the idea of revising and updating the MRC framework. It was further developed at a workshop co-convened with Sally Macintyre and Janet Darbyshire, and organised with the help of Linda Morris on 15-16 May 2006. Workshop participants and others with an interest in the evaluation of complex interventions were invited to comment on a draft of the revised guidance, which was also reviewed by members of the MRC Health Services and Public Health Research Board and MRC Methodology Research Panel. A full list of all those who contributed suggestions is provided in the full guidance document.

Funding: MRC Health Services and Public Health Research Board and the MRC Population Health Sciences Research Network.

Competing interests: None declared.

Provenance and peer review: Not commissioned; externally peer reviewed.

1 Medical Research Council. *A framework for the development and evaluation of RCTs for complex interventions to improve health* . London: MRC, 2000.
2 Campbell M, Fitzpatrick R, Haines A, Kinmonth AL, Sandercock P, Spiegelhalter D, et al. Framework for the design and evaluation of complex interventions to improve health. *BMJ* 2000;321:694-6.
3 Hardeman W, Sutton S, Griffin S, Johnston M, White A, Wareham NJ, et al. A causal modelling approach to the development of theory-based behaviour change programmes for trial evaluation. *Health Educ Res* 2005;20:676-87.
4 Campbell NC, Murray E, Darbyshire J, Emery J, Farmer A, Griffiths F, et al. Designing and evaluating complex interventions to improve health care. *BMJ* 2007;334:455-9.
5 Oakley A, Strange V, Bonell C, Allen E, Stephenson J, Ripple Study Team. Process evaluation in randomised controlled trials of complex interventions. *BMJ* 2006;332:413-6.
6 Campbell M, Donner A, Klar N. Developments in cluster randomised trials and Statistics in Medicine. *Stat Med* 2007;26:2-19.
7 Shiell A, Hawe P, Gold L. Complex interventions or complex systems? Implications for health economic evaluation. *BMJ* 2008;336:1281-3.
8 Belsky J, Melhuish E, Barnes J, Leyland AH, Romaniuk H, National Evaluation of Sure Start Research Team. Effects of Sure Start local programmes on children and families: early findings from a quasi-experimental, cross sectional study. *BMJ* 2006;332:1476-81.
9 Haynes B. Can it work? Does it work? Is it worth it? The testing of healthcare interventions is evolving. *BMJ* 1999;319:652-3.
10 Michie S, Abraham C. Interventions to change health behaviours: evidence-based or evidence-inspired? *Psychol Health* 2004;19:29-49.
11 Muir H. Let science rule: the rational way to run societies. *New Scientist* 2008;198:40-3.
12 Creegan C, Hedges A. *Towards a policy evaluation service: developing infrastructure to support the use of experimental and quasi-experimental methods* . London: Ministry of Justice, 2007.
13 Thomson H, Hoskins R, Petticrew M, Ogilvie D, Craig N, Quinn T, et al. Evaluating the health effects of social interventions. *BMJ* 2004;328:282-5.
14 Ogilvie D, Mitchell R, Mutrie N, Petticrew M, Platt S. Evaluating health effects of transport interventions: methodologic case study. *Am J Prev Med* 2006;31:118-26.

15 Academy of Medical Sciences. *Identifying the environmental causes of disease: how should we decide what to believe and when to take action* ? London: Academy of Medical Sciences, 2007.
16 Torgerson D, Byford S. Economic modelling before clinical trials. *BMJ* 2002;325:98.
17 Eldridge S, Spencer A, Cryer C, Pearsons S, Underwood M, Feder G. Why modelling a complex intervention is an important precursor to trial design: lessons from studying an intervention to reduce falls-related injuries in elderly people. *J Health Services Res Policy* 2005;10:133-42.
18 Eldridge S, Ashby D, Feder G, Rudnicka AR, Ukoumunne OC. Lessons for cluster randomized trials in the twenty-first century: a systematic review of trials in primary care. *Clin Trials* 2004;1:80-90.
19 Power R, Langhaug L, Nyamurera T, Wilson D, Bassett M, Cowan F. Developing complex interventions for rigorous evaluation—a case study from rural Zimbabwe. *Health Educ Res* 2004;19:570-5.
20 Howden-Chapman P, Crane J, Matheson A, Viggers H, Cunningham M, Blakely T, et al. Retrofitting houses with insulation to reduce health inequalities: aims and methods of a clustered community-based trial. *Soc Sci Med* 2005;61:2600-10.
21 Black N. Why we need observational studies to evaluate the effectiveness of health care. *BMJ* 1996;312:1215-8.
22 Glasziou P, Chalmers I, Rawlins M, McCulloch P. When are randomised trials unnecessary? Picking signal from noise. *BMJ* 2007;334:349-51.
23 MacMahon S, Collins R. Reliable assessment of the effects of treatment on mortality and major morbidity. II. observational studies. *Lancet* 2001;357:455-62.
24 Gunnell D, Fernando R, Hewagama M, Priyangika W, Konradsen F, Eddleston M. The impact of pesticide regulations on suicide in Sri Lanka. *Int J Epidemiol* 2007;36:1235-42.
25 Clancy L, Goodman P, Sinclair H, Dockery DW. Effect of air pollution control on death rates in Dublin, Ireland: an intervention study. *Lancet* 2002;360:1210-4.
26 Haw SJ, Gruer L. Changes in exposure of adult non-smokers to secondhand smoke after implementation of smoke-free legislation in Scotland: national cross sectional survey. *BMJ* 2007;335:549-52.
27 Wortman PM. An exemplary evaluation of a program that worked: the High/Scope Perry preschool project. *Am J Eval* 1995;16:257-65.
28 Nazareth I, Freemantle N, Duggan C, Mason J, Haines A. Evaluation of a complex intervention for changing professional behaviour: the evidence based out reach (EBOR) trial. *J Health Services Res Policy* 2002;7:230-8.
29 Freemantle N, Nazareth I, Eccles M, Wood J, Haines A, EBOR Triallists. A randomised controlled trial of the effect of educational outreach by community pharmacists on prescribing in UK general practice. *Br J Gen Pract* 2002;52:290-5.
30 Hawe P, Shiell A, Riley T. Complex interventions: how “out of control” can a randomised trial be? *BMJ* 2004;328:1561-3.
31 Farmer A, Wade A, Goyder E, Yudkin P, French D, Craven A, et al. Impact of self-monitoring of blood glucose in the management of patients with non-insulin treated diabetes: open parallel group randomised trial. *BMJ* 2007;335:132-9.
32 Patton G, Bond L, Butler H, Glover S. Changing schools, changing health? Design and implementation of the Gatehouse Project. *J Adolesc Health* 2003;33:231-9.
33 Abraham C, Michie S. A taxonomy of behavior change techniques used in interventions. *Health Psychol* 2008;27:379-87.
34 Glasziou P, Meats E, Heneghan C, Shepperd S. What is missing from descriptions of treatments in trials and reviews? *BMJ* 2008;336:1472-4.
35 Boutron I, Moher D, Altman D, Scultz K, Ravaud P. Extending the CONSORT statement to randomized trials of non-pharmacologic treatment: explanation and elaboration. *Ann Int Med* 2008;148:295-309.
36 Von Elm E, Altman D, Egger M, Pocock SJ, Gotzsche P, Vandenbroucke JP, et al. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ* 2007;335:806-8.

Accepted: 23 August 2008

Cite this as: *BMJ* 2008;337:a1655

Figure

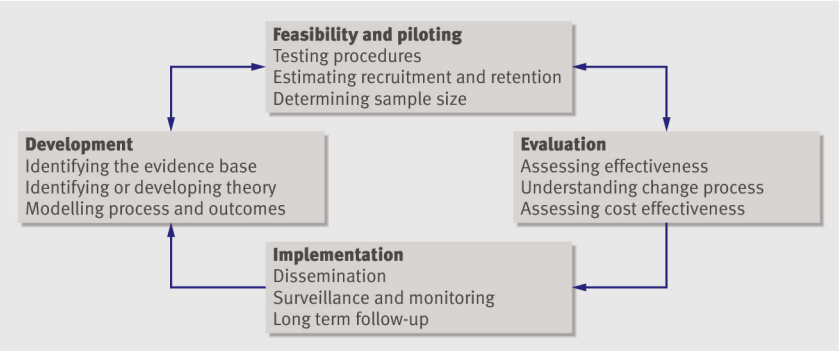


Fig 1 Key elements of the development and evaluation process

death rates—that are strikingly similar to those seen in England. And time series show few, if any, dramatic changes in trends as a result of reforms or investment. So what has the massive investment in quality initiatives bought? Was it worth it? And are there any new levers available to pull?

Bevan's verdict?

How would Bevan rate this performance against his founding principles? On universality, he would certainly be content. But on equity and quality he would be far from happy. For age, sex, socioeconomic group, and geography it's easy to uncover glaring inequities of access and use. Rather than providing services of world beating quality, there's enough comparative data from similar countries to judge the NHS's outcomes of care as distinctly average (or worse).[12] It's as if most of Bevan's successors had simply forgotten that equity and quality were founding principles of the NHS. Next week, I will be considering whether the founding principle of comprehensiveness has fared any better.

Tony Delamothe deputy editor, *BMJ*, London WC1H 9JR
tdelamothe@bmj.com

In the preparation of this article I have greatly benefited from discussions with Sheila Leatherman, Julian Le Grand, Veena Raleigh, and Julian Sheather.

Competing interests: None declared.

1 Ministry of Health, Department of Health for Scotland. *A national health service*. London: HMSO, 1944. (Cmd 6502.)
2 Department of Health. *Failed asylum seekers and ordinary residence—advice to overseas visitors managers*. 2008. www.dh.co.uk (search for: 9854).
3 Dixon A, Le Grand J, Henderson J, Murray R, Poteliakhoff E. *Is the NHS equitable? A review of the evidence*. London: London School of Economics, 2003. (Health and social care discussion paper No 1.)
4 Pell JP, Pell ACH, Norrie J, Ford I, Cobbe SM. Effect of socioeconomic deprivation on waiting time for cardiac surgery: retrospective cohort study. *BMJ* 2000;320:15-8.
5 Lawson N. *Machines, markets and morals: the new politics of a democratic NHS*. London: Compass, 2008.
6 Le Grand J. *The strategy of equality: redistribution and the social services*. London: George Allen & Unwin, 1982.
7 Department of Health. *Tackling inequalities: a programme for action*. London: DoH, 2003.
8 Klein R. *The new politics of the NHS: from creation to reinvention*. Abingdon: Radcliffe, 2006.
9 Department of Health. *Tackling health inequalities: 2007 status report on the programme for action*. London: DoH, 2008.
10 Gubb J. Why the NHS is the sick man of Europe. *Civitas Rev* 2008;5(1):1-11.
11 Department of Health. *Our NHS, our future. NHS next stage review interim report*. London: DoH, 2007.
12 Leatherman S, Sutherland K. *The quest for quality: refining the NHS reforms*. London: Nuffield Trust, 2008.
13 Kmietowicz Z. Plan to end age discrimination in NHS is launched. *BMJ* 2001;322:751.
14 Eaton L. Help the Aged accuses NHS of discrimination. *BMJ* 2002;324:564.
15 Young J. Ageism in services for transient ischaemic attack and stroke. *BMJ* 2006;333:508-9.
16 Doyal L. Sex, gender, and health: the need for a new approach. *BMJ* 2001;323:1061-3.
17 Hippisley-Cox J, Pringle M, Crown N, Meal A, Wynn A. Sex inequalities in ischaemic heart disease in general practice: cross sectional survey. *BMJ* 2001;322:832.
18 Gill PS, Kai J, Bhopal RS, Wild S. *Black and minority ethnic groups*. www.hcna.bham.ac.uk/series/bemgframe.htm
19 Parliamentary Office of Science and Technology. Ethnicity and health. *Postnote* 2007;276:1. www.parliament.uk/documents/upload/postpn276.pdf
20 Hitchen L. Compulsory recording of patients' ethnic data will help show disease trends. *BMJ* 2008;336:1039.
21 Sekhri N, Timmis A, Chen R, Junghans C, Walsh N, Zaman J, et al. Inequity of access to investigation and effect on clinical outcomes: prognostic study of coronary angiography for suspected stable angina pectoris. *BMJ* 2008;336:1058-61.
22 Leatherman S, Sutherland K. *The quest for quality: a mid-term evaluation of the ten-year quality agenda guide*. London: Nuffield Trust, 2003.

Complex interventions or complex systems? Implications for health economic evaluation

Although guidelines exist for evaluating complex interventions, they may be of little help in dealing with the multiple effects of interventions in complex systems such as hospitals. **Alan Shiell**, **Penelope Hawe**, and **Lisa Gold** explain why it is important to distinguish the two types of complexity

Health researchers commonly use the notion of complexity to indicate the problems faced in evaluating the effectiveness of many non-drug interventions.[1-3] However, although it is rarely delineated, complexity has two meanings. In the first it is a property of the intervention, and in the second it is a property of the system in which the intervention is implemented. We examine the implications of these two views for economic evaluation.

What do we mean by complex?

The first view of complexity, in effect, means complicated. This is the meaning used in the Medical Research Council's framework for the evaluation of complex interventions.[4 5] A complex intervention is "built up from a number of components, which may act both independently and inter-dependently."[4] This makes it hard to define the "active ingredients" and to be sure which component or combinations of components is more important.

The second view makes reference to the insights offered by complexity science.[6-9] Complexity is a property of a system not an intervention. A complex system is one that is adaptive to changes in its local environment, is composed of other complex systems (for example, the human body), and behaves in a non-linear fashion (change in outcome is not proportional to change in input).

Complex systems include primary care, hospitals, and schools. Interventions in these settings may be simple or complicated, but the complex systems approach makes us consider the wider ramifications of intervening and to be aware of the interaction that occurs between components of the intervention as well as between the intervention and the context in which it is implemented. This includes the operations, structures, and relations that exist in each setting[10 11] and the implications that contextual effects have for designing and evaluating interventions.[12 13]

The distinction between the two approaches (complex interventions versus complex systems) is easily blurred because they share common features—for example, non-standardisation, multiplicity, interactions. Analysts working with complex interventions, for example, also recognise the importance of context.[14] Furthermore, complicated interventions can take on the characteristics of complex systems, since it is impossible to separate the intervention from the human agency required for its delivery.[15] However, it is important to recognise the differences between the two approaches and to identify when each one is being applied correctly when thinking about economic evaluation.

Implications for economic evaluation

The main challenge in evaluating complex interventions arises because the active elements of the intervention are subject to more variation than in typical drug trials. Campbell and colleagues,[5] citing the operation of a stroke unit, point to variation among units in staff characteristics, clinical practices, manage-

ment protocols, and infrastructure. This makes it difficult to specify what the intervention is, what is most effective, or how to replicate the intervention beyond the original trial.[5]

When it comes to economic evaluation, however, the problem of specifying the intervention is less of an issue. Economists compare the value of what goes in (the resources) with what comes out (the outcomes). If you can specify the inputs and outcomes with sufficient clarity to ensure that changes in resource use and benefits can be measured and valued, then it is not necessary to understand how the intervention works.

To illustrate this we can use the example of group psychotherapy.[5] This is a complex intervention because the content of each consultation is tailored to the individual needs of the patients in the group and is adapted as the programme of consultations unfolds and each client responds in their different ways to treatment. We do not know whether the treatment effect is the result of the content of the consultation, the personal style of the therapist, the dynamic of the group, a combination of all three, or some other consideration. However, to evaluate economic efficiency the economist need only know how long each session lasts, how many sessions there are, how many people are in each group (so that costs can be apportioned to patients), what skill level is required of the therapist (so that salary costs can be generalised), what other resources are required (the venue for example), and what effect treatment has on health outcomes and use of health services. The content of the consultation is immaterial.

Of course, economic evaluation of multicomponent interventions does present challenges. It is more difficult to draw boundaries around the evaluation. Multicomponent interventions to reduce excessive alcohol consumption, for example, will benefit people beyond the problem drinker, including family members and the community at large, which raises questions about how to include such benefits in the appraisal. But simple interventions tackling the same problem also generate these externalities. Multicomponent interventions will also have an effect on multiple dimensions of health and have non-health benefits as well, but then so too do many simple interventions (vaccination being a good example).

PEFKOS/FOTOLIA

Since everything is connected, changes in one part of a complex system feed through to other parts

Thus, complex interventions of the sort discussed by the MRC are more difficult to evaluate, but there is nothing substantively different about their economic evaluation. No new economic methods are required, and the problems can all be solved with time, effort, and resources.[16]

In contrast, evaluating the economic efficiency of interventions directed at changing the properties of complex systems presents big challenges. Complex systems have several defining characteristics including the tendency to be self organising, be sensitive to initial conditions, and make non-linear phase transitions (to jump quickly from one position to another very different position); the existence of emergent properties; and the importance of interaction effects and feedback.[17] These characteristics affect what measures of effectiveness should be included in the economic evaluation and how the consequences of the intervention are valued.

What should we evaluate?

The economist's concern with value will always mean looking for improvement in final (health) outcomes. However, the characteristics of complex systems suggest the need to do much more than this.

Firstly, evaluation of outcomes typically involves measuring health changes at the individual level and simply summing these to capture the "social" effect. In a complex system this is no longer wholly appropriate. Complex systems have emergent properties that are a feature of the system as a whole.[18] These properties are not seen in any one part of a complex system nor are they summations of individual parts (community empowerment,[19] social exclusion, and income inequality are noted emergent properties relevant to population health). So outcomes should be measured at multiple levels within the complex system, with tools designed specifically for this purpose.

Secondly, the relatively short follow-up periods of most intervention studies and the fact that non-linear change in complex systems is difficult to observe in its early stages means there is high risk of missing important outcomes and concluding prematurely that the intervention is not effective. The impact of public health advocacy on public health policy such as gun control is a case in point. Multiple "advocacy episodes" may have no discernible impact on policy, but then a tipping point is reached, a phase transition occurs, and new laws are introduced. In the search for cause and effect, the role played by advocates in creating the conditions for change is easily overlooked in favour of prominent and immediately prior events.[20] To minimise the risk of premature evaluation and wrongful attribution, economists must become comfortable working with evidence of intermediate changes in either process or impact that act as preconditions for a phase transition.

One important indicator of system level change is movement in the positions of key actors within the structures that make up the complex system, and with it changes in their relationships with other actors and agencies. Relational data (collected at the individual level but analysed at the network level using social network methods) are needed to capture these effects. In community based interventions to improve access to primary care, for example, we might wish to see family practitioners become more influential in the network of providers. In interventions in schools, a reduction in the number of children or teachers who are socially isolated, and corresponding increases in the density of support networks, might provide evidence of effect.[21] Such organisational and social network measures are not final outcomes favoured by economists (the economic test depends on whether such changes lead to improvements in health and wellbeing) but they provide reassuring evidence that the intervention is having an effect on the system, which will in turn hopefully lead to improvements in health. We are beginning to see these network analytical methods introduced into cluster randomised controlled trials.[22]

How should we evaluate benefits?

The consequences of intervention in a complex system will not be the small scale, marginal changes usually examined by economists. Since everything is interconnected, changes in one part of the system feed through to other parts of the system and feedback on themselves. The economist's usual approach assumes that the

effects of the intervention can be examined in isolation of changes in the broader context. Feedback loops are ignored. With interventions in complex systems, this no longer applies. Nothing can be assumed constant as everything is linked to everything else.

Two consequences follow. Firstly, spin-off effects are to be expected. The consequences of system level change are both multiple and multiplied, with induced costs and outcomes beyond those originally envisaged in the research protocol. The practical challenge of identifying and capturing these effects within an evaluation is substantial.[22]

Secondly, one of the things that economists assume is unchanging is the value (that is, the importance) that people assign to the intervention. This assumption is unlikely to hold with system level change. We see this most notably in tobacco control, where the concerted action of public health advocates to reduce the harm associated with tobacco use has changed behaviours and social norms. Support for banning smoking in public places often increases after the policy is implemented.[23] This means that the value of an intervention that changes the dynamic of a complex system is likely to be a function of that intervention: people value the intervention more after implementation than before it. Preferences are no longer stable, and this undermines the validity of the methods economists use to ascertain value. More collective, deliberative methods of eliciting social value are needed.[24]

New approaches?

The view that complexity refers to the systems in which interventions are implemented affects all efforts to evaluate interventions, not just those of economists. For example, it is difficult to attribute causality in a complex system, not least because such systems are sensitive to initial conditions and miniscule differences at baseline can lead to very large differences in outcome. Thus, randomisation (even at the cluster level) may not eliminate all causes of bias, even if it removes all observable differences between groups.[25]

The economic evaluation of interventions aimed at changing systems requires new ways of thinking: one sensitive to ecological theory, interactions between microlevel and macrolevel variables, nonlinearities, multiplier effects, and the fact that individual values are shaped by the interventions we seek to evaluate and the contexts we seek to change.

SUMMARY POINTS

Health research often uses complex to refer to multicomponent interventions

An alternate view is that complexity refers to systems

Interventions implemented in complex systems are likely to have diverse, far-reaching, and nonlinear effects

Distinguishing the two types of complexity is important for economic evaluation

The methodological agenda is huge, and the proper evaluation of systems level change will be expensive. We should remember therefore that existing methods have served us relatively well thus far. Linear approximation may be sufficient to assess non-linear change (and it is easier and less expensive). Our concerns do not rule out the use of current economic approaches. They do, however, point to the need for extensive prospective data collection alongside cluster trials to capture signs of non-linear change, unintended consequences, and multiplier effects,[13] and for more extensive use of modelling to assess the sensitivity of economic evaluations to the inclusion of these effects.

We need to recognise whether we have a complex intervention or an intervention in a complex system, and whether the dynamic characteristics of the system matter enough for us to change our evaluation approach. Neither question is easy to answer, making efforts to develop the means of diagnosing complexity especially important.

Alan Shiell professor, Population Health Intervention Research Centre, University of Calgary, Calgary T2N 4N1, Canada
ashiell@ucalgary.ca

Penelope Hawe professor, Population Health Intervention Research Centre, University of Calgary, Calgary T2N 4N1, Canada

Lisa Gold senior research fellow, Health Economics Unit, School of Health and Social Development, Deakin University, Melbourne, Australia

We thank colleagues in the International Collaboration on Complex Interventions funded by the Canadian Institutes of Health Research (CIHR) and colleagues at an MRC workshop convened to consider revisions to the original MRC Framework. We also thank the reviewers for helpful suggestions. The views expressed here are those of the authors alone.

Contributors and sources: PH is the Markin chair in health and society. AS holds a Canadian Institutes of Health Research chair in the economics of public health. Both are also health scientists funded by the Alberta Heritage Foundation for Medical Research. AS had the original idea to write the paper based on ideas developed in discussion with PH and LG. AS wrote the first draft and revisions. PH and LG provided comments on early drafts, helped to refine the arguments and suggested edits. AS is guarantor.

Competing interests: None declared.

Provenance and peer review: Not commissioned; externally peer reviewed.

Accepted: 15 April 2008

1 Byrne M, Cupples ME, Smith SM, Leathem C, Corrigan M, Byrne MC, et al. Development of a complex intervention for secondary prevention of coronary heart disease in primary care using the UK Medical Research Council framework. *Am J Manage Care* 2006;12:261-6.
2 Oakley A, Strange V, Bonell C, Allen E, Stephenson J. Process evaluation in randomised controlled trials of complex interventions. *BMJ* 2006;332:413-6.
3 Wolff N. Randomised trials of socially complex interventions: promise or peril? *J Health Serv Res Policy* 2001;6:123-6.
4 Medical Research Council. *A framework for the development and evaluation of randomised controlled trials for complex interventions to improve health*. London: MRC, 2000.
5 Campbell M, Fitzpatrick R, Haines A, Kinmouth AL, Sandercock P, Spiegelhalter D, et al. Framework for design and evaluation of complex interventions to improve health. *BMJ* 2000;321:694-6.
6 Begun JW, Zimmerman B, Dooley KJ. Health care organizations as complex adaptive systems. In: Mick SS, Wyttenback ME, eds. *Advances in health care organization theory*. San Francisco: Jossey-Bass, 2003:253-88.
7 Miller WL, McDaniel RR, Crabtree BF, Stange KC. Practice jazz: understanding variation in family practices using complexity science. *J Fam Pract* 2001;50:872-9.
8 Plsek PE, Greenhalgh T. The challenge of complexity in health care. *BMJ* 2001;323:625-8.
9 Zimmerman B, Lindberg C, Plsek P. *Edgeware: insights from complexity science for health care leaders*. Irving, TX: Veterans Health Administration, 1998.
10 Goodwin MA, Zysanski SJ, Zronek S, Ruhe M, Wever SM, Konrad N, et al. A clinical trial of tailored office systems for preventive service delivery: the study to enhance prevention by understanding practice (STEP-UP). *Am J Prev Med* 2001;21:20-8.
11 Leischow SJ, Milstein B. Systems thinking and modelling for public health practice. *Am J Public Health* 2006;96:403-5.
12 Litaker D, Tomolo A, Liberatore V, Stang KC, Aron D. Using complexity theory to build interventions that improve health care delivery in primary care. *J Gen Intern Med* 2006;21:S30-4.
13 Hawe P, Shiell A, Riley T. Complex interventions: how out of control can a randomised controlled trial be? *BMJ* 2004;328:1561-3.
14 Campbell NC, Murray E, Darbyshire J, Emery J, Farmer A, Griffiths F, et al. Designing and evaluating complex interventions to improve health care. *BMJ* 2007;334:455-9.
15 Pawson R, Greenhalgh T, Harvey G, Walshe K. Realist review—a new method of systematic review designed for complex policy interventions. *J Health Serv Res Policy* 2005;10:21-34.
16 Byford S, Sefton T. Economic evaluation of complex health and social interventions. *National Institute Economic Review* 2003;186:98-108.
17 Rickles D, Hawe P, Shiell A. A simple guide to chaos and complexity. *J Epidemiol Community Health* 2007;61:93-7.
18 Halley JD, Winkler DA. Classification of emergence and its relation to self organisation. *Complexity* 2008 Feb 27 doi: 10.1002/cplx.20216.
19 Zimmerman MA. Taking aim on empowerment research: on the distinction between individual and psychological conceptions. *Am J Community Psychol* 1984;18:169-77.
20 Chapman S. Advocacy in public health: roles and challenges. *Int J Epidemiol* 2001;30:1226-32.
21 Hawe P, Ghali L. Use of social network analysis to map the social relationships of staff and teachers at school. *Health Educ Res* 2008;23:62-9.
22 Hawe P, Shiell A, Riley T, Gold L. Methods for exploring implementation variation and local context within a cluster randomised community intervention trial. *J Epidemiol Community Health* 2004;58:788-93.
23 Siahpush M, Scollo M. Trends in public support for smoking bans in public places in Australia. *Austr N Z J Public Health* 2001;25:473.
24 Sagoff M. *Policy analysis and social values*. In: Carrow MM, Churchill RP, Cordes JJ, eds. *Democracy, social values, and public policy*. Westport, CT: Greenwood Publishers, 1998:91-106.
25 Rickles D. Causality in complex interventions. *Medicine, Health Care, and Philosophy* (in press).

Complex interventions: how "out of control" can a randomised controlled trial be?

Penelope Hawe, Alan Shiell, Therese Riley

Complex interventions are more than the sum of their parts, and interventions need to be better theorised to reflect this

Many people think that standardisation and randomised controlled trials go hand in hand. Having an intervention look the same as possible in different places is thought to be paramount. But this may be why some community interventions have had weak effects. We propose a radical departure from the way large scale interventions are typically conceptualised. This could liberate interventions to be responsive to local context and potentially more effective while still allowing meaningful evaluation in controlled designs. The key lies in looking past the simple elements of a system to embrace complex system functions and processes.

Divergent views

The suitability of cluster randomised trials for evaluating interventions directed at whole communities or organisations remains vexed.[1] It need not be.[2] Some health promotion advocates (including the WHO European working group on health promotion evaluation) believe randomised controlled trials are inappropriate because of the perceived requirement for interventions in different sites to be standardised or look the same.[1 3 4] They have abandoned randomised trials because they think context level adaptation, which is essential for interventions to work, is precluded by trial designs. An example of context level adaptation might be adjusting educational materials to suit various local learning styles and literacy levels.

Lead thinkers in complex interventions, such as the UK's Medical Research Council, also think that trials of complex interventions must "consistently provide as close to the same intervention as possible" by "standardising the content and delivery of the intervention."[5] By contrast, however, they do not see this as a reason to reject randomised controlled trials.

These divergent views have led to problems on two fronts. Firstly, the field of health promotion is being turned away from randomised controlled trials.[1 3 4] This could have heavy consequences for the future accumulation of high quality evidence about prevention. Secondly, when trials with organisations and whole communities do go ahead, the story is consistently becoming one of expensive failure—that is, weak or non-significant findings at huge cost.[6–8] Could one of the reasons for the interventions not working be that the components have been overly standardised?

Something has to change. The current view about standardisation is at odds with the notion of complex systems. We believe that an alternative way to view standardisation could allow state of the art interventions (and ones that might look different in different sites) to be more effective and to be meaningfully evaluated in a randomised controlled trial. First, however, we have to re-examine our understanding of the term complex intervention.

What is a complex intervention?

The MRC document *A Framework for the Development and Evaluation of Randomised Controlled Trials for Complex Interventions* argues that "the greater the difficulty in defining precisely what exactly are the 'active ingredients' of an intervention and how they relate to each other, the greater the likelihood that you are dealing with a *complex* intervention."[5] The document gives examples of complex interventions from the setting up of new healthcare teams, to interventions to get treatment guidelines adopted, to whole community education interventions. Setting aside the problem that this definition is also consistent with a poorly thought through intervention, we believe that the field could benefit by delving further into complexity science.

Complexity is defined as "a scientific theory which asserts that some systems display behavioral phenomena that are completely inexplicable by any conventional analysis of the systems' constituent parts."[9] Reducing a complex system to its component parts amounts to "irretrievable loss of what makes it a system."[9] Those of us who have decomposed interventions into components for process evaluation might feel uncomfortable at this point. Yes, we may have been able to describe an intervention, say, simply in terms of the percentage of general practitioners who attend the training workshops and the percentage of patients who report having read the leaflets. Thinking about process evaluation in this way is the norm.[10 11] But by doing so, have we really captured the essence of the intervention? We have, if all we think our intervention to be is the sum of the parts. But that is not, by definition, a complex intervention. It remains a simple one.

Standardising complex interventions

So, could a controlled trial design (which requires something to be replicable and recognisable as the intervention in each site) ever be appropriate to evalu-

Department of Community Health Sciences, University of Calgary, 3330 Hospital Drive NW, Calgary T2N 4N1, Alberta, Canada
Penelope Hawe *professor*
Alan Shiell *professor*
Therese Riley *postdoctoral fellow*
Correspondence to: P Hawe
phawe@ucalgary.ca

BMJ 2004;328:1561–3

ate a (truly) complex intervention? The answer is yes. The crucial point lies in "what" is standardised. Rather than defining the components of the intervention as standard—for example, the information kit, the counselling intervention, the workshops—what should be defined as standard are the steps in the change process that the elements are purporting to facilitate or the key functions that they are meant to have. For example, "workshops for general practitioners" are better regarded as mechanisms to engage general practitioners in organisational change or train them in a particular skill. These mechanisms could then take on different forms according to local context, while achieving the same objective. [12] (table).

Defining integrity of interventions

With most (simple) interventions, integrity is defined as having the "dose" delivered at an optimal level and in the same way in each site.[10] Complex intervention thinking defines integrity of interventions differently. The issue is to allow the form to be adapted while standardising the process and function. Some precedents exist here. For example, Mullen and colleagues conducted a meta-analysis of 500 patient education trials and showed that interventions were more likely to be effective if they met particular criteria fitting with behavioural change theory—for example, being tailored to the patient's individual learning needs or being set up to provide feedback about a patient's progress.[17] The indicators of quality were driven by theory and concerned the functions provided by the key elements of the intervention rather than the elements themselves (such as a video).

Context level adaptation does not have to mean that the integrity of what is being evaluated across multiple sites is lost. Integrity defined functionally, rather than compositionally, is the key.

Real world contexts

We are not the first to think this way. In school health, Durlak discussed non-standard interventions that "cannot be compartmentalised into a predetermined number and sequence of activities."[18] This sounds like complex interventions. Characterised by activities like capacity building and organisational change, these interventions have specific, theory driven principles that ensure that non-standard interventions (different forms in different contexts) conform to standard processes. They are still evaluable by randomised controlled trials. Indeed, a randomised controlled trial of such an intervention (which is "out of control" to some ways of thinking) might be exactly what is required to provide more convincing evidence that community development interventions are effective.

More studies of this type would help to reverse the current evidence imbalance when policy makers weigh up "best buys" in health promotion. At present they often have to compare traditional areas like asthma education (which usually come with randomised controlled trial evidence) with community development (which is usually supported only with case study evidence).[19] The more conservative, patient targeted interventions backed by randomised controlled trials generally win hands down.[19]

Rethinking ways to use the intervention-context interaction to maximum effect may make complex interventions stronger. The MRC document on complex intervention trials calls for standardisation but also recognises the need in the exploratory phase to "describe the constant and variable components of a replicable intervention."[5] But it does not say how to make this distinction.

An alternative way of thinking about standardisation may help. The fixed aspects of the intervention are the essential functions. The variable aspect is their form in different contexts. In this way an intervention evaluated in a pragmatic, effectiveness, or real world trial would not be defined haphazardly, as it sometimes is now,[20] as the default option for whenever researchers were not able to accomplish the standardised components that they idealised. Instead, with lateral thinking, theorising about the real world context would become the ideal,[21 22]reversing current custom.[23] That is, instead of mimicking trial phases which assume that the "best" or the "ideal" comes from the laboratory and gets progressively compromised in real world applications, community trial design would start by trying to understand communities themselves as complex systems and how the health problem or phenomena of interest is recurrently produced by that system.

Example of alternative ways to standardise a whole community intervention to prevent depression in a cluster trial*

Principle of intervention	Type of standardisation	
	By form	By function
To educate patients about depression	All sites distribute the same written patient information kit	All sites devise ways to distribute information tailored to local literacy, language, culture, and learning styles
To improve detection, management, and referral of patients in primary care	All sites hold a series of three in-service training workshops for general practitioners with preset curriculums	Local health authorities are provided with materials and resources to devise in-service training tailored to local schedules, venues, and preferred learning methods
To involve local residents and decision makers in order to increase uptake, effectiveness, and sustainability of the intervention	A local intervention steering committee is convened in each site with representatives of pre-specified organisations	Mechanisms are devised to engage local key agencies and consumers in decision making about the intervention. Suggested options: steering committee, consultations, surveys, website, phone-ins
To harness and facilitate material, emotional, informational, and affirmational support across social networks of people in particular life stages	All mothers of new babies are invited to join discussion and mutual support groups. People moving into nursing homes receive three friendly visits from a designated resident	Methods to alter network size, network diversity, contact frequency, reciprocity, or types of exchanges are tailored to subgroup preferences

* Hypothetical example drawing on published studies[13-16] and reflecting a sample of principles depending on the intervention theory.

Summary points

Standardisation has been taken to mean that all the components of an intervention are the same in different sites

This definition treats a potentially complex intervention as a simple one

In complex interventions, the function and process of the intervention should be standardised not the components themselves

This allows the form to be tailored to local conditions and could improve effectiveness

Intervention integrity would be defined as evidence of fit with the theory or principles of the hypothesised change process

Conclusion

The shackles of simple intervention thinking may prove hard to throw off. Although an intervention may be described as complex, the signs of simple intervention thinking will be apparent in how the intervention is described and whether integrity is tied to the extent to which certain standardised forms are present. Investigators should justify the approach they take with interventions—that is, whether interventions are theorised as simple or complex. Complex systems rhetoric should not become an excuse to mean "anything goes." More critical interrogation of intervention logic may build stronger, more effective interventions.

Contributors and sources: All authors were collaborators in a cluster randomised intervention trial in maternal health promotion.[14] All are participating in a newly funded international collaboration on complex interventions funded by the Canadian Institutes of Health Research. PH drafted the original idea for the paper based on experience and conversations with TR and AS. All contributed to developing the idea and writing the paper.

Funding: PH and AS are senior scholars of the Alberta Heritage Foundation for Medical Research. PH is also supported by an endowment as Markin Chair in Health and Society at the University of Calgary.

Competing interests: None declared.

1 Nutbeam D. Evaluating health promotion-progress, problems and solutions. *Health Promotion Int* 1998;13:27-44.
2 Oakley A. *Experiments in knowing.* Cambridge: Polity Press, 2000.
3 Tones K. Evaluating health promotion: a tale of three errors. *Patient Educ Counsel* 2000;39:227-36.
4 World Health Organisation Europe. *Health promotion evaluation: recommendations for policy makers.* Copenhagen: WHO Working Group on Health Promotion Evaluation, 1999.
5 Medical Research Council. *A framework for the development and evaluation of randomised controlled trials for complex interventions to improve health.* London: MRC, 2000.
6 Secker-Walker RH, Gnich W, Platt S, Lancaster T. Community interventions for reducing smoking among adults. *Cochrane Database Syst Rev* 2002:3;CD001745.
7 Susser M. The tribulations of trials. *Am J Public Health* 1995;85:156-8.
8 Thompson B, Coronado G, Snipes SA, Puschel K. Methodological advances and ongoing challenges in designing community based health promotion interventions. *Annu Rev Public Health* 2003;24:315-40.
9 Casti JL. *Would-be worlds: how simulation is changing the frontiers of science.* New York: John Wiley, 1997.
10 Flora JA, Lefebvre RC, Murray DM, Stone EJ, Assaf A, Mittelmark MB, et al. A community education monitoring system: methods from the Stanford Five-City Project, the Minnesota Heart Health Intervention and the Pawtucket Heart Health Intervention. *Health Educ Res* 1993;8:81-95.
11 Hawe P, Degeling D, Hall J. *Evaluating health promotion: a health workers guide.* Sydney: MacLennan and Petty, 1990.
12 Castro FG, Barrera M, Martinez CR. The cultural adaptation of preventive interventions: resolving tensions between fidelity and fit. *Prev Sci* 2004;5:41-5.
13 Llewellyn-Jones RH, Baikie KA, Smithers H, Cohen J, Snowden J, Tennant CC. Multifaceted shared care intervention for later life depression in residential care: a randomised trial. *BMJ* 1999;319:677-82.
14 Lumley J, Small R, Brown S, Watson L, Gunn J, Mitchell C, and Dawson W. PRISM (program of resources, information and support for mothers) protocol for a community-randomised trial. *BMC Public Health* 2003;3:36.
15 Paton J, Jenkins R, Scott J. Collective approaches for the control of depression in England. *Soc Psychiatry Psychiatric Epidemiol* 2001;36:423-8.
16 Israel BA. Social networks and social support: implications for natural helping and community level interventions. *Health Educ Q* 1985;12: 65-80.
17 Mullen PD, Green LW, Persinger GS. Clinical trials of patient education for chronic disease: a comparative meta analysis of intervention types. *Prev Med* 1985;14:735-81.
18 Durlak JA. Why intervention implementation is important. *J Prev Intervent Community* 1998;17(2):5-18.
19 Hawe P, Shiell A. Preserving innovation under increasing accountability pressures: the health promotion investment portfolio approach. *Health Promot J Aust* 1995;5(2):4-9.
20 McMahon AD. Study control, violators, inclusion criteria and defining explanatory and pragmatic trials. *Stat Med* 2002;21:1365-76.
21 Bauman LJ, Stein REK, Ireys HT. Reinventing fidelity: the transfer of social technology among settings. *Am J Community Psychol* 1991;19: 619-39.
22 Ottoson JM, Green LW. Reconciling concept and context: theory of implementation *Adv Health Educ Promot* 1987;2:353-82.
23 Flay BR. Efficacy and effectiveness trials (and other phases of research) in the development of health promotion interventions. *Prev Med* 1986;15:451-74.

(Accepted 24 March 2004)

Birth of a baby girl and social stigma

While working as a junior resident in India, I was posted to the neonatology ward of a hospital serving a rural area, where most of the babies born belonged to families from the surrounding countryside.

I soon realised that the birth of a baby girl was regarded as a calamity by the family, particularly by the father's mother. It was considered so bad that sometimes even the mother detested her newborn baby (although emotionally still cuddling her). The mother, still recovering from the trauma of the delivery, fearfully anticipated the possibility of rejection by her in-laws. In the worst cases the poor baby girl was abandoned by the family and left for adoption. In contrast, if a baby boy was born it was a joyous occasion. The family would bring sweets for the nurses and doctors as a mark of happiness and gratitude. I was really shaken by seeing this level of discrimination faced by baby girls.

Then it happened, a baby girl was born and we all got sweets. The family was overjoyed with the news of the birth of the baby girl. This came as a surprise to all of the hospital staff. Later on, I learnt from one of the nursing staff that the baby was the first girl child in this family after two generations. Then I thought that all was not lost and a silver lining could be seen in the grey clouds.

I wish that every baby girl born in this world could receive a similar welcome. Since then I have cherished this dream that one day this social stigma of having a baby girl will disappear from our society.

Afshan Salim *paediatrician, Hull*

TRIALS

Theory of Change: a theory-driven approach to enhance the Medical Research Council's framework for complex interventions

De Silva *et al.*

BioMed Central

De Silva *et al. Trials* 2014, **15**:267
http://www.trialsjournal.com/content/15/1/267

De Silva et al. Trials 2014, 15:267
http://www.trialsjournal.com/content/15/1/267

TRIALS

METHODOLOGY Open Access

Theory of Change: a theory-driven approach to enhance the Medical Research Council's framework for complex interventions

Mary J De Silva[1*], Erica Breuer[2], Lucy Lee[1], Laura Asher[1], Neerja Chowdhary[3], Crick Lund[2] and Vikram Patel[1,3]

Abstract

Background: The Medical Research Councils' framework for complex interventions has been criticized for not including theory-driven approaches to evaluation. Although the framework does include broad guidance on the use of theory, it contains little practical guidance for implementers and there have been calls to develop a more comprehensive approach. A prospective, theory-driven process of intervention design and evaluation is required to develop complex healthcare interventions which are more likely to be effective, sustainable and scalable.

Methods: We propose a theory-driven approach to the design and evaluation of complex interventions by adapting and integrating a programmatic design and evaluation tool, Theory of Change (ToC), into the MRC framework for complex interventions. We provide a guide to what ToC is, how to construct one, and how to integrate its use into research projects seeking to design, implement and evaluate complex interventions using the MRC framework. We test this approach by using ToC within two randomized controlled trials and one non-randomized evaluation of complex interventions.

Results: Our application of ToC in three research projects has shown that ToC can strengthen key stages of the MRC framework. It can aid the development of interventions by providing a framework for enhanced stakeholder engagement and by explicitly designing an intervention that is embedded in the local context. For the feasibility and piloting stage, ToC enables the systematic identification of knowledge gaps to generate research questions that strengthen intervention design. ToC may improve the evaluation of interventions by providing a comprehensive set of indicators to evaluate all stages of the causal pathway through which an intervention achieves impact, combining evaluations of intervention effectiveness with detailed process evaluations into one theoretical framework.

Conclusions: Incorporating a ToC approach into the MRC framework holds promise for improving the design and evaluation of complex interventions, thereby increasing the likelihood that the intervention will be ultimately effective, sustainable and scalable. We urge researchers developing and evaluating complex interventions to consider using this approach, to evaluate its usefulness and to build an evidence base to further refine the methodology.

Trial registration: Clinical trials.gov: NCT02160249

Keywords: Complex interventions, Theory of Change, MRC framework for complex interventions

* Correspondence: mary.desilva@lshtm.ac.uk
[1]Centre for Global Mental Health, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK
Full list of author information is available at the end of the article

Background

The updated Medical Research Council (MRC) framework for complex interventions [1] is a set of guidelines for designing and evaluating complex interventions which has been widely influential in the field [2]. The framework emphasizes four phases of intervention development, feasibility and piloting, evaluation, and implementation which take place as an iterative rather than a linear process. However, the MRC framework has been criticized for not including theory-driven approaches to evaluation [3]. Although the framework does reference theory-driven approaches, it does not explicitly recommend any, or provide guidance on how to incorporate them into the design and evaluation of complex interventions [1]. The evaluation of complex interventions has also been criticized for not providing a clear explanation of the mechanisms of change through which the intervention leads to real-world impact, and for not examining how the intervention interacts with context [4]. These omissions reflect the paucity of practical examples of the use of theory-driven approaches that have been shown to work, resulting in calls for researchers to provide such examples so that the MRC framework can reflect current best practice [2,3,5].

In order to develop complex interventions which are more likely to be effective, sustainable and scalable, evaluators need to understand not just whether, but how and why an intervention has a particular effect, and which parts of a complex intervention have the greatest impact on outcomes. For this, a prospective, theory-driven process of intervention design and evaluation is required.

In this article we propose a theory-driven approach to the design and evaluation of complex interventions by adapting and integrating an existing approach, Theory of Change (ToC), into the MRC framework. We provide a guide to what ToC is, how to construct one, and how to integrate its use into research projects seeking to design, implement and evaluate complex interventions using the MRC framework.

What is Theory of Change?

Theory-driven approaches to program evaluation can be traced back to the 1930s [6], with further development by among others Kirkpatrick in the late 1950s [7] and Chen in the 1980s [8]. Their basic tenet is that understanding the theory underlying a program approach is necessary to understand whether, and how, it works [6]. ToC developed organically, influenced by program evaluation theorists, theories of social change [9] and the work of the Aspen Institute Roundtable on Community Change in the 1990s [10-12]. This organic development has resulted in no standardized definition of ToC [13]. We will refer to ToC as that developed by the Aspen Institute and promoted by organizations such as ActKnowledge, who set up the Centre for Theory of Change and support capacity building in its use[a].

ToC is 'a theory of how and why an initiative works' [10] which can be empirically tested by measuring indicators for every expected step on the hypothesized causal pathway to impact. It is developed in collaboration with stakeholders and modified throughout the intervention development and evaluation process through an 'ongoing process of reflection to explore change and how it happens' [9]. It is visually represented in a ToC map which is a graphic representation of the causal pathways through which an intervention is expected to achieve its impact within the constraints of the setting in which it is implemented (see Figure 1 for an example).

ToC has been used to design and evaluate development programs in many different contexts globally [14-18]. Recognizing its capacity to provide a framework for monitoring, evaluation and learning throughout a program cycle [11], ToC is increasingly being used by international donors such as the Gates Foundation, the UK Department for International Development (DfID), Comic Relief and Grand Challenges Canada, to monitor and evaluate their research and development programs [9,13].

ToC is not a sociological or psychological theory such as Complexity Theory [19] or the Theory of Planned Behaviour [20], but a pragmatic framework which describes how the intervention affects change. The ToC can be strengthened by inserting sociological or psychological theories at key points to explain why particular links happen. For example, behavioral change theories may explain why community awareness-raising activities increase uptake of services as one link in a ToC describing how to improve maternal and child health outcomes. Equally, a ToC approach is complimentary to other frameworks which seek to reduce the chance of implementation failure, such as Normalization Process Theory (NPT) [21]. While NPT provides a framework detailing what questions should be asked to design an intervention that is more likely to be 'normalized' into routine practice, ToC provides an explanation for how these questions can be answered. ToC can also be used to strengthen randomized controlled trials (RCTs) and other evaluations by building and validating program theories of interventions that are then empirically tested [4].

Although similar to other theory-driven approaches to evaluation, ToC differs in a number of key ways. Logic models, for example, present a simplified model of action in a rigid linear way which articulates inputs, activities and outcomes but which does not make explicit how they are linked, or measure whether they have been achieved. Logical frameworks (log frames) are also rigidly structured and include resources, inputs, outputs, outcomes, impacts and assumptions, as well as indicators

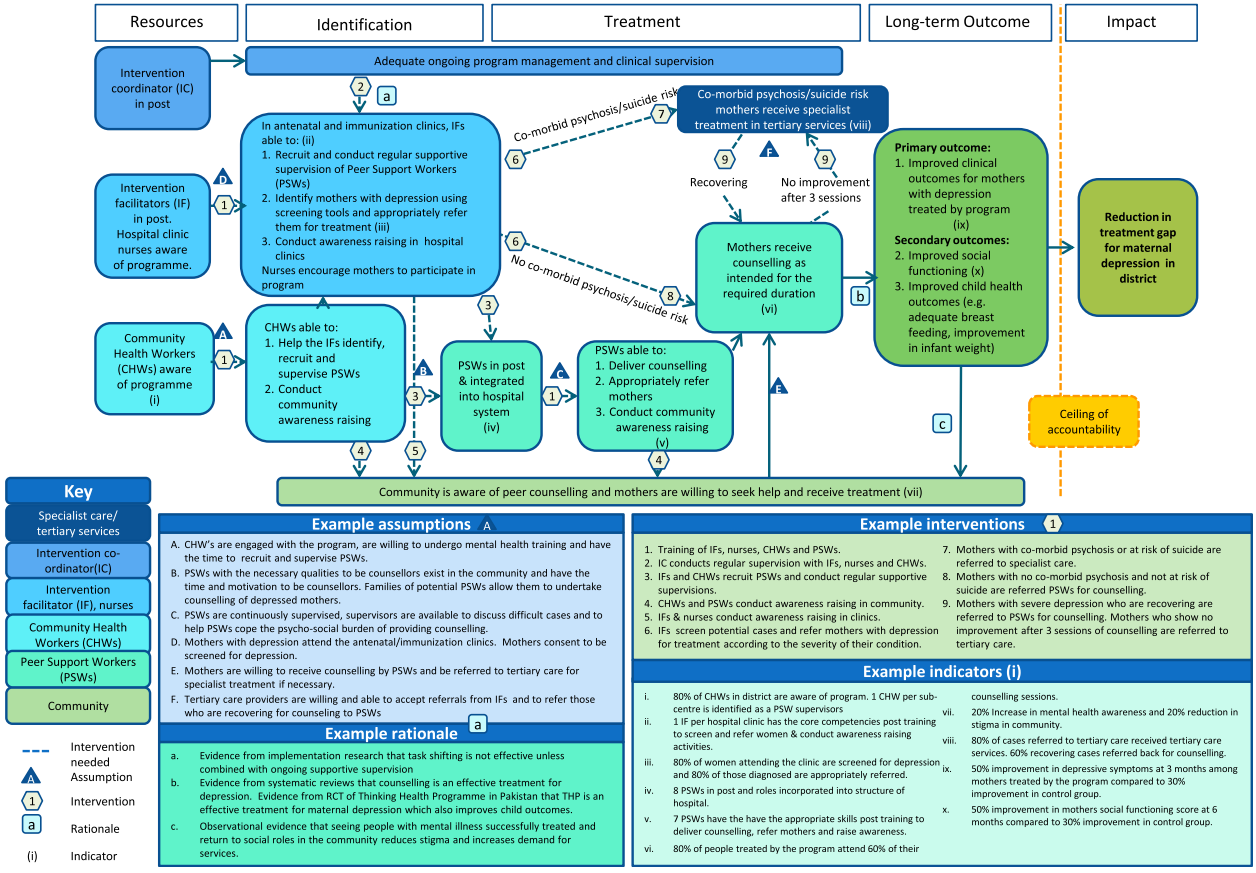


Figure 1 SHARE Theory of Change: peer counselling for maternal depression in Goa, India.

for success and specific milestones to measure. However, log frames do not necessarily explain how the various components work together in a causal pathway to achieve the impact [22], and do not link activities to outcomes. Although suitable for program monitoring and evaluation, these approaches are less useful in a research setting where the understanding of the mechanisms underlying the intervention is a key goal in unpacking the 'black box' of complex health interventions.

ToC has a number of advantages over these approaches. Firstly, ToC is a more flexible format which makes explicit the causal pathways through which the outcomes and activities work to achieve the desired impact, but which does not impose a pre-defined structure (such as linear structures in logic models or a cycle as in project cycle management) [23]. Instead, ToC allows for multiple causal pathways, levels of interventions and feedback loops which better reflect the reality of how complex interventions achieve their impact. Secondly, the articulation of the evidence base as the rationale for each link (pre-condition) in the causal pathway ensures that each step along the causal pathway is evidence based. Lastly, as the achievement of each pre-condition is measured through an indicator, this allows for a detailed understanding of how and whether an intervention is working and which components of a complex intervention are the most important in achieving impact.

Although ToC has been used in a research context, it is not a well-known approach in evaluation methods for complex health interventions. In a systematic review in preparation, we found 51 articles which used ToC to some extent in the design, implementation or evaluation of public health interventions (E Breuer, personal communication). However, most did not use ToC systematically throughout the research process or did not describe in significant detail how the ToC informed the development or evaluation of their intervention. None of the papers reported using ToC in RCTs or suggested using ToC together with the MRC framework.

Methods

We are currently piloting the use of ToC to design, implement and evaluate complex interventions for mental health in a number of research projects in low- and middle-income countries. These include both RCTs and observational designs to which ToC is also suited. Throughout the

paper we use the example of the South Asian Hub for Advocacy, Research and Education on mental health (SHARE) trial [b] to illustrate the process of developing a ToC within the MRC framework. SHARE is adapting an evidence-based counselling intervention for maternal depression delivered by Community Health Workers in Pakistan [24] to be delivered by peer support workers as this is more sustainable in a low resource context. The effectiveness of the peer-delivery system is being evaluated through a cluster RCT in Pakistan and an individual RCT in India. The SHARE example also demonstrates that ToC can be used both to develop new interventions and also to adapt existing interventions to new contexts or models of service delivery. To provide further examples, Case Study 1 describes the use of ToC in the Rehabilitation Intervention for people with Schizophrenia in Ethiopia (RISE) trial, and Case Study 2 describes the use of ToC in a non-randomized evaluation in the PRogramme for Improving Mental health care (PRIME), integrating mental health into primary care in five low- and middle-income countries.

Ethical approval

Ethical approval for SHARE, including the ToC workshops, was granted by the Indian Council of Medical Research, Sangath Institutional Review Board, India, and the London School of Hygiene and Tropical Medicine, UK (reference 7141). Ethical approval for RISE was including the ToC workshops was granted by the Addis Ababa University College of Health Sciences Institutional Review Board (reference 039/13/PSY), the Addis Ababa University Department of Psychiatry (reference MF/PSY/212/2005) and from the London School of Hygiene and Tropical Medicine, UK (reference 6408). Ethical Approval for PRIME was granted by the University of Cape Town (reference HREC 412/2011) and from Institutional Review Boards in each of the five participating countries, as well as by the World Health Organization. Either verbal or written informed consent was obtained from all of the participants in the ToC workshops in all the projects.

Results

The results describe how ToC was applied to each phase of the MRC framework (development, piloting, evaluation and dissemination) in the context of the SHARE trial. The two case studies provide further practical examples of how ToC can be used in combination with each stage of the MRC framework to develop and evaluate complex interventions.

Development of complex interventions using Theory of Change

At the start of the intervention development phase, ToC uses a participatory approach by bringing together a range of stakeholders (for example health service planners, healthcare workers and service users) to develop a ToC map and to encourage stakeholder buy-in to the project [25]. This takes the form of a series of workshops, interviews or focus groups, with the choice of method based upon what is locally feasible and acceptable [15].

In the workshop, stakeholders first agree on the real-world impact they want to achieve. They then identify the causal pathways through which this change can be achieved in that context using the available resources. These are articulated as a series of preconditions leading to outcomes, the order of which can be adjusted as the pathway develops. Determining what contextual conditions are necessary to achieve the outcomes, what resources are required to implement the interventions, and how the program gains the commitment of those resources are crucial outputs of the process. There are several guidelines available which may assist with conducting ToC workshops [12,26].

Additional components of the ToC map include: identifying the interventions needed to move from one precondition on the causal pathway to the next and articulating the evidence for each link in the pathway. This rationale may be drawn from a range of sources including research evidence, behaviour change theories, local knowledge or from primary research conducted as part of the intervention feasibility and piloting stage. Drawing on a more diverse set of evidence and experience should produce a more plausible intervention. In addition, the key assumptions which set out the conditions which the causal pathway needs to achieve impact are highlighted. Through this process, potential barriers and interventions needed to overcome these barriers can be identified so that the ultimate impact can be achieved. Lastly, indicators are identified for each precondition in the pathway to evaluate whether each stage of the pathway leading to the final impact is achieved.

All these components are displayed graphically on a ToC map, often with an accompanying narrative that describes the pathways and key assumptions. Figure 1 presents the ToC map for SHARE India and Table 1 elaborates on common ToC terminology and definitions outlined above.

In SHARE, the research team who developed the original intervention in Pakistan constructed a ToC map describing how the intervention worked. This was used as the basis of ToC workshops in India to modify the intervention to be delivered by peer support workers, adapt it to the Indian context and facilitate stakeholder buy-in to the project. Eighteen health professionals (9 doctors, 3 gynecologists and 2 psychiatrists) and 11 other professionals (3 counsellors, 5 staff nurses and 3 community maternal health workers) participated in a half-day ToC workshop held in the district hospital where the trial was to be conducted, facilitated by the research

Table 1 Common Theory of Change terminology and definitions

Terminology	Definition	Examples
Impact (ultimate outcome, goal)	The real-world change you are trying to affect. The program may contribute towards achieving this impact, and not achieve it solely on its own.	- Reduced prevalence of depression in a district.
Longterm outcome	The final outcome the program is able to change on its own. This will be the primary outcome of the evaluation.	- Reduced symptoms of depression in the population receiving the intervention
Precondition (short-term, intermediate and longterm outcomes, milestones)	The intended results of the interventions. Things that don't exist now, but need to exist in order for the logical causal pathway not to be broken and the impact achieved. The logical and sequential connections between shorter-term preconditions and longer-term outcomes that are illustrated on the ToC diagram as arrows.	- Staff in post to develop intervention. - Changes in knowledge, attitudes and skills of health workers to enable them to successfully deliver the intervention.
Ceiling of accountability	Level at which you stop using indicators to measure whether the outcomes have been achieved and therefore stop accepting responsibility for achieving those outcomes. The ceiling of accountability is often drawn between the impact and the longterm outcome.	- Project aims to change individual patient outcomes, but does not accept responsibility for changing levels of health problems in the wider population (the goal), as it cannot achieve this on its own (though it may contribute to this wider goal).
Indicator	Things you can measure and document to determine whether you are making progress towards, or have achieved, each outcome.	-Number of staff trained - Knowledge of and attitudes towards mental illness among carers - Percentage of people with mental illness diagnosed in primary care - Reduction in clinical severity of mental illness
Interventions (strategies)	The different components of the complex intervention. A dotted arrow is used to show when an intervention is needed to move from one outcome to the next. A solid arrow is used when one outcome logically leads to the next without the need for any intervention.	- Training program for service providers - Community awareness campaign - Inter-personal therapy - Antidepressant medication
Rationale	Key beliefs that underlie why one outcome is an outcome for the next, and why you must do certain activities to produce the desired outcome. Can be based on evidence or experience.	- Mothers and their families need to be educated about the signs and symptoms of maternal depression in order for maternal depression to be detected in the community.
Assumptions	An external condition beyond the control of the project that must exist for the outcome to be achieved.	- Political desire to support the program exists - Funder continues to fund project - Task-sharing is politically and culturally acceptable

team. The output from the SHARE workshop comprised a ToC map (Figure 1) and a detailed report generated from an analysis of the group discussions outlining the barriers in delivering the intervention and strategies to overcome them.

Feasibility and piloting complex interventions using Theory of Change

Before an intervention is implemented, the ToC should be tested in the feasibility and piloting phase of the MRC framework. This involves using assumptions articulated in the ToC to formulate research questions to test in formative research. This may help reduce implementation failure as weak links in the causal pathway are tested and strengthened, leading to a revision of the intervention where necessary. The ToC is then modified to reflect changes resulting from the feasibility and piloting phase and a revised ToC is taken forward for formal testing in the evaluation phase. Developing a ToC must be a continual process of reflection and adaptation as barriers to implementation arise and new evidence comes to light, requiring pathways to be changed and strengthened.

The assumptions generated by SHARE'S ToC were used to generate questions to be tested in the intervention's formative research. Key assumptions being tested through qualitative interviews with community members and mothers include 'peer support workers with the necessary qualities to be counsellors exist in the community and have the time and motivation to be counsellors' (Figure 1, assumption B), and 'mothers are willing to receive counselling by peer support workers' (Figure 1, assumption E). Other formative research methods to test key assumptions include an analysis of patient flow through the antenatal and immunization clinics where mothers with depression will be identified, an assessment of the existing referral system for specialist mental health care, and qualitative interviews with clinic staff to determine the most acceptable and feasible methods of screening mothers attending the clinics.

Evaluating complex interventions using Theory of Change

The evaluation stage of a complex intervention using a ToC approach involves identifying at least one indicator for every precondition within that framework to measure whether it has been achieved. Indicators must be specific enough to describe what change is necessary in the precondition to move up the causal pathway (for example how many people need to be trained in order to deliver the intervention as intended). Pre-specifying the level of change needed to achieve an precondition makes it easier to design the components of the intervention to achieve that target. It also ensures that the indicators are meaningful measures of whether a precondition has been achieved or not. For example in SHARE, we measure whether the peer support workers have acquired the skills from training in order to deliver the counselling as intended, rather than simply recording how many people have been trained.

Evaluation using a ToC framework involves measuring indicators at all stages of implementation, not just an intervention's primary and secondary outcomes. This includes a wider range of input, process, output and outcome indicators than may normally be measured, with a clear focus on measuring whether key stages in the causal pathway are achieved. ToC can therefore be used as the theoretical framework on which to base a detailed process evaluation necessary to unpack the 'black box' of a complex intervention [5,27]. ToC allows for multiple outcomes of the intervention to be pre-specified within a theoretical framework, thereby explicitly evaluating the multiple outcomes that complex interventions may lead to. In SHARE, multiple preconditions to be captured by the evaluation include the core competencies of peer support workers, the willingness of mothers with depression to seek and receive treatment, as well as the long-term outcomes of the impact of the intervention on maternal clinical, social and economic outcomes, as well as on child health.

As a result, an evaluation based on ToC will require a number of different methods to capture all of the indicators. In SHARE, the evaluation includes an RCT to assess the effect of peer-counselling on patient outcomes, nested studies of the fidelity of training including an assessment of the competencies achieved by peer support workers and the quality of supervision received, and collection of clinic based data to measure key preconditions in the ToC map such as the proportion of women who are referred to peer-counselling who receive treatment, and their adherence to the sessions.

The analysis of data collected using a ToC approach has the potential to combine process and effectiveness indicators into a single analysis which can help untangle whether, how and why an intervention has an impact in a particular context, and whether it may be suitable for scale-up or for adaptation to new settings. In order for this to be achieved, appropriate modelling techniques need to be applied, drawing on methods from other fields such as structural equation modelling [28], discrete-event simulation models [29], agent-based modelling [30], and system dynamics modelling [31]. The application of these methods to the analysis of complex interventions is an important area for further research.

Implementing complex interventions using Theory of Change

Experience of implementation and evidence gathered from the evaluation is combined to revise the ToC and produce the final 'story' of how the intervention worked in a particular setting. This provides a comprehensive description of the intervention which can be disseminated to a variety of audiences, providing information on the components of the intervention that need to be adapted for use in other settings. The MRC guidance calls for more detailed and standardized descriptions of complex interventions in published reports to facilitate exchange of knowledge and to encourage synthesis of results from similar studies [1,32]. As the projects described in this paper are still ongoing, it remains to be tested whether ToC is a useful tool to meet this challenge. A full description of Case Study 1 and Case Study 2 can be found below.

Case Study 1 | Use of Theory of Change in the RISE Trial

Background

The RISE trial (Rehabilitation Intervention for people with Schizophrenia in Ethiopia) aims to develop and test in a cluster-randomized trial, community-based rehabilitation (CBR) for people with schizophrenia in Sodo, a rural district in Ethiopia. CBR is a multi-sectoral method for improving social inclusion and functioning in people

with disabilities [33]. CBR has been shown to improve outcomes in people with schizophrenia in India [34], but intervention development work was needed to design an intervention suitable for Ethiopia, a setting with fewer public sector resources. A situational analysis, literature review and review of existing CBR guidelines and projects were undertaken first. This allowed us to identify potential CBR components for RISE, including health (for example adherence support), social (for example social skills training), livelihood (for example, support returning to work), empowerment (self-help groups) and education (literacy group) elements.

Development of the intervention

Two ToC workshops were held with key stakeholders to determine the feasibility of delivering these intervention components in Sodo district. The first half-day workshop involved eight national experts in CBR and mental health. The second half-day workshop was held in Sodo and included 20 community leaders, including district-level representatives of microfinance, education, police, traditional healers and religious leaders. The ToC map was created at the first workshop and presented to and refined in the second workshop. Additional file 1 lists a summary version of the ToC map. Through these workshops, the CBR components were finalized and the key delivery structures were developed. For example, the key decision was made that CBR should be delivered by CBR workers, specially recruited and trained for RISE, rather than existing government community health workers. The workshops also allowed us to recognize the richness of local resources, and how these might be utilized for CBR, for example literacy groups and edirs (burial associations).

Feasibility and piloting of the intervention

Following the ToC workshop, we conducted 16 qualitative interviews and five focus groups with people with schizophrenia, caregivers, community leaders, existing CBR workers (for people with physical disabilities), and community and primary healthcare workers to test the assumptions identified in the ToC map. For example, a key concern was that it would be difficult to find and retain local CBR fieldworkers willing to work with people with schizophrenia, due to concerns about safety and stigma. The qualitative interviews showed that if adequate safety and supervision mechanisms were provided (for example risk assessment) recruitment and retention would be possible. A second assumption, that community leaders would be willing to participate without personal gain, generated conflicting views from different stakeholder groups. Female caregivers, based on their previous experiences, were skeptical that community leaders would provide support, whilst community leaders themselves were keen to collaborate. These differing opinions highlighted the importance of the pilot in understanding how CBR will work in practice. The ToC map was amended using the qualitative results and will continue to be adapted following the pilot, which will be conducted in mid-2014.

Evaluation of the intervention

The preconditions, long-term outcomes and indicators arising from the ToC map were used to plan a comprehensive and meaningful evaluation for RISE which combines an assessment of both the effectiveness of the intervention and also the process of implementation. One strength of CBR is that it is tailored to individual needs, meaning each CBR recipient receives a different 'version' of CBR. However, this means it is difficult to evaluate which CBR component, or synergy between components, results in positive outcomes for recipients. Using ToC allowed us to conceptualize how different CBR components fit onto the causal pathway to improved functioning in people with schizophrenia, and to develop appropriate ways to evaluate each component. Ultimately this will allow us to determine the active ingredients of CBR and how the process of implementation affects outcomes, in order to adapt and refine the intervention for scaling up in Ethiopia, or to translate it for implementation in new settings.

Challenges

A challenge of using ToC was the difficulty in operationalizing true ownership of the ToC map by stakeholders in the workshops. Although stakeholders provided the content, the map itself was created and 'owned' by the researchers throughout the process. This may have been due to the short time frame for explaining the concepts behind both ToC and CBR, before asking for participation in creating the map.

Case study 2 | Use of Theory of Change in the PRrogramme for Improving Mental health care (PRIME)

Background

PRIME is developing and evaluating district level mental health care plans integrating mental health services into primary care in five low- and middle-income countries (India, South Africa, Ethiopia, Uganda and Nepal) [35]. Within PRIME, we used ToC as a conceptual framework underpinning the development and evaluation of the mental health care plans at a country level and also at a cross-country level to provide a framework highlighting commonalities across all five countries. The use of ToC in the PRIME program is described in detail elsewhere [36].

Development of the intervention

The PRIME Cross Country ToC was developed with 15 members of the PRIME team from all countries at a

workshop in Goa, India at the start of the program. This initial ToC described the causal pathways of how the PRIME interventions would need to work in order to achieve the ultimate impact of 'improved health, social, and economic outcomes for people with priority disorders and their families/carers in the PRIME districts'. A summary version of the PRIME Cross Country ToC is shown in Additional file 2.

Following the drafting of the cross-country ToC, individual countries developed district specific ToCs during a series of ToC workshops which are described in detail elsewhere [36]. In brief, between two and four workshops were held in each country with stakeholders including policymakers, district level health planners and management, mental health specialists, researchers, and service providers. The size of the workshops varied significantly between countries with a median of 15 (interquartile range 13 to 22) stakeholders attending each workshop. The workshops provided an opportunity to develop logical, evidence-based ToC maps with stakeholders, contextualize the mental health care plans and elicit buy-in from stakeholders and acted as a forum for knowledge exchange between researchers and stakeholders. Stakeholders provided detailed knowledge on the functioning of the health system and information about local resources which could be mobilized for the implementation of the mental health care plans. The researchers provided guidance on the development on the ToC, the evidence available for potential interventions, as well as strategies to evaluate the success of the plans. The ToC maps were further developed after the ToC workshops and used as a basis for the development of the district specific mental health care plans, in combination with a variety of other methods including a situational analysis [37], a costing tool, and interviews and focus group discussions with key stakeholders. The Cross Country ToC was then further refined by comparing it to the district specific ToC maps to ensure that all the key preconditions and long-term outcomes across countries were captured.

Feasibility and piloting of the intervention

The cross-country ToC highlighted a number of assumptions which were used to develop cross-country topic guides for formative semi-structured interview guides and focus group discussions with stakeholders designed to test the feasibility of the interventions. These were supplemented by questions designed to answer country-specific assumptions taken from the district level ToCs. The subsequent qualitative interviews and focus groups gathered information in each country on access and demand for mental health care, service delivery recovery and rehabilitation and accountability. The results of this formative research were used to refine the district specific ToCs and develop the mental health care plans in each country.

Evaluation of the intervention

The indicators for the cross-country ToC were refined using the indicators from the district specific ToC maps and compared across countries to identify common indicators across countries that could be used as the basis of an evaluation strategy to answer cross-country research questions such as whether the mental health care plans reduce the treatment gap in the districts, and whether the patients treated by the programs have improved clinical, social and economic functioning. These indicators were used to plan the evaluation design for PRIME. A variety of evaluation methodologies are being used, including detailed process evaluations, repeat cross-sectional surveys, cohort studies and RCTs. As the PRIME evaluation is ongoing, we have not yet been able to test whether the process and outcome indicators from the ToC can be combined in a single analysis or to test the usefulness of ToC in the implementation of the interventions at scale. These will be the subject of future research by PRIME.

Challenges

One of the challenges in PRIME was using multiple ToC maps at different levels. The PRIME cross-country ToC map provided us with an overall framework of the causal pathways required for the integration of mental health care into primary health care but did not specify the country specific context and resources. In particular, the interventions which will be implemented in each country as part of the mental health care plan are different for each district according to local feasibility, existing financial and human resources and cultural acceptability. For this reason, a locally adapted district level ToC was essential for each country to ensure that these factors are accounted for in the design and evaluation of their mental health care plan. However, having an overarching ToC allowed a cross-country view of how the programs were likely to work in all countries which led to the development of an evaluation design which could be used across all countries. Another limitation of the ToC approach is that if it is to be developed with stakeholders, it requires a significant amount of work facilitating the ToC workshops and compiling the resulting ToC. However, as this process is structured around the components of the ToC and has a defined output, it is an efficient way to conduct discussions with stakeholders [36]. Critical to the success of ToC in PRIME has been having a 'ToC champion' who took responsibility for coordinating with countries to help them develop their district level ToC, and drove forward the development and refinement of the cross-country ToC.

Discussion

Our experience of using ToC in three projects designing and evaluating complex interventions to improve mental

health has demonstrated a number of benefits, which we believe strengthen the existing MRC framework. Figure 2 summarizes how using ToC has the potential to strengthen each phase in the MRC framework.

Using a ToC approach for the development of an intervention may enhance the MRC framework in two key ways. Firstly, using a ToC approach provides a useful framework to guide stakeholder engagement. While stakeholder participation is an increasingly an important part of health services research[c], using a ToC approach may prompt a deeper level of engagement than other methods as it enables stakeholders to take part in the initial design of the intervention in a formal and participatory way. This was certainly true in the PRIME and RISE projects where we found a deeper level of stakeholder engagement from a relatively short workshop. However, our experience from all three projects indicates that this stakeholder engagement in the ToC process does not extend beyond the workshops, and that a ToC champion within the project is needed to drive the process forward.

Secondly, it may improve the initial design and potential effectiveness of the intervention by explicitly designing interventions which are embedded in the local context and seek to have an impact in the real world as opposed to in a research setting. Designing a feasible intervention that is likely to work in the constraints of the context and available resources is challenging. Agreeing on how interventions lead to outcomes can be politically charged if achieving those outcomes implies a major resource reallocation, or changes in work patterns away from the current status. One of the strengths of ToC is that design and implementation issues are brought centre-stage from the start, and if any aspects of the intervention are politically unacceptable, or if the resources will not be available, then all stakeholders have to compromise and come to alternative solutions to ensure that the impact is achieved. This was demonstrated in the RISE trial where very early on in the workshop it became clear that using government community health workers to deliver the intervention as we had planned would not be politically acceptable, leading the group to decide to train dedicated CBR workers.

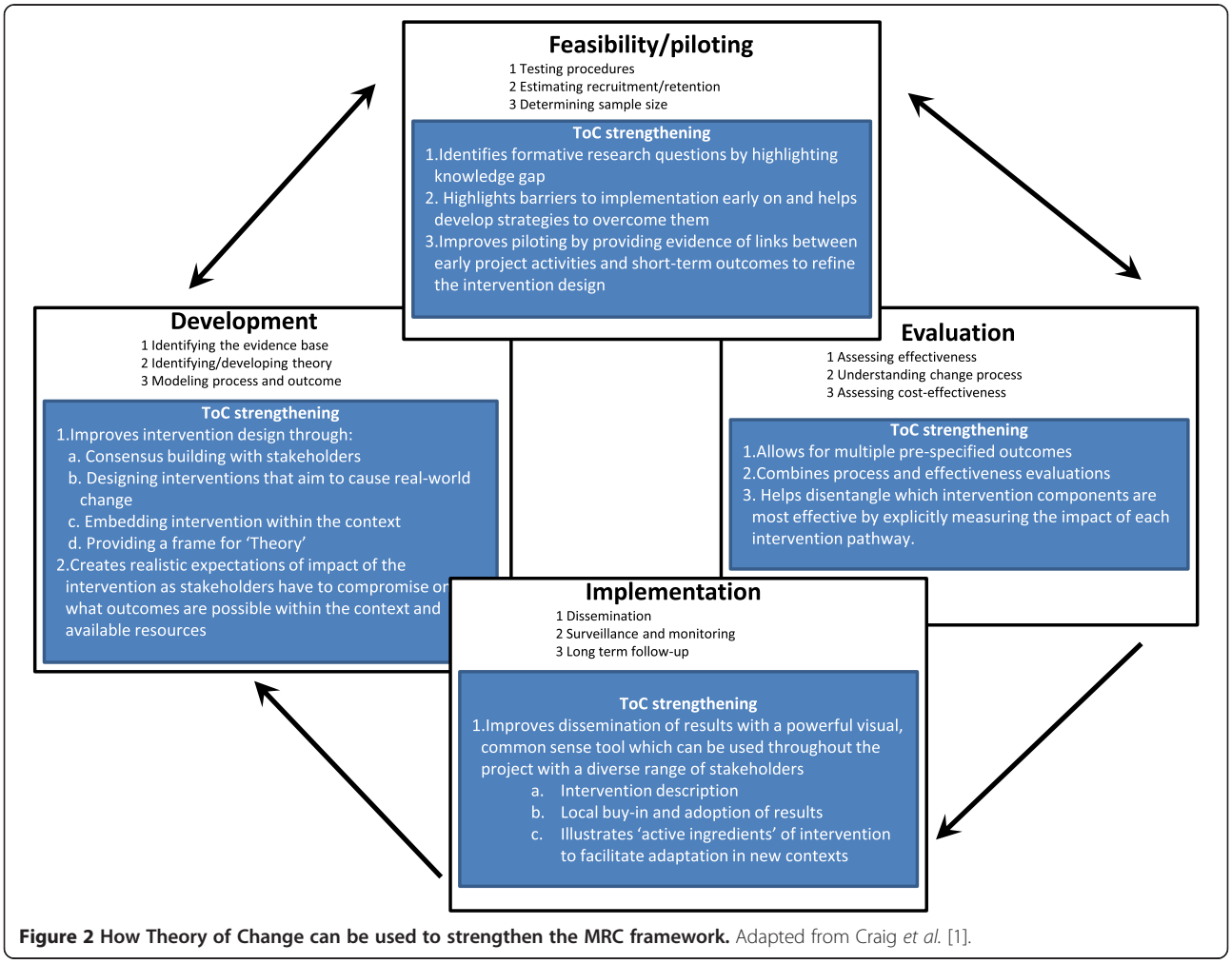


Figure 2 How Theory of Change can be used to strengthen the MRC framework. Adapted from Craig *et al.* [1].

Another advantage of the ToC process is embedding the intervention within the context in which it is to be implemented, which enables contextual factors which may affect implementation to be highlighted and tracked, along with potential unintended consequences of the intervention. This was also shown in RISE where the workshops highlighted the richness of local resources that could be utilized for the CBR intervention, such as local literacy groups and burial associations that the CBR workers could refer people to. In PRIME, we have designed district, primary healthcare facility and community level case studies to track changes in the local context (such as changes in local health priorities or staffing levels in primary healthcare facilities) that may affect the impact that the mental health care plans have. By forcing us to measure not only the process of implementing the interventions but also the context in which it is implemented, we hope to be able to conduct a much richer analysis of how and why the PRIME mental health care plans achieve any impact. This is particularly important in evaluations of complex interventions where the context may facilitate or impede the success of the intervention [38].

One key advantage of using ToC to pilot the feasibility of interventions is that it enables the systematic identification of knowledge gaps to generate research questions for the pilot stage. Completing the rationale for each link in the causal pathway highlights which linkages lack evidence and therefore what additional work is needed to fill those gaps. Secondly, highlighting specific barriers to intervention delivery early on enables strategies to overcome these barriers to be incorporated into the intervention design. An example of this from SHARE is the need for consensus building workshops with policymakers and hospital staff to change attitudes towards using peer-counsellors for treating maternal depression, which we have now made part of the intervention.

A key intended benefit of using a ToC framework for the evaluation of complex interventions, particularly in trials, is that it breaks down the barriers between evaluations of intervention effectiveness and process evaluations by combining them into one framework. Though detailed process evaluations are becoming more widely used in trials, they are rarely combined with an assessment of intervention effectiveness in a single analysis, enabling interpretation of the outcome data in light of the process data. As the three projects we describe in the paper have not yet reached the analysis stage, it remains unknown whether this benefit will be realized. Future work needs to explore ways of modelling the pathways to impact by combining process and outcome data, enabling a more nuanced assessment of which components of the intervention may be most critical for achieving the desired outcome.

Our research has shown that ToC is useful in the implementation phase of the MRC framework as it helps to develop locally adapted, contextually relevant plans developed with stakeholders, including local policymakers, which are therefore more likely to be feasible and acceptable and work within existing resource constraints. ToC may confer important benefits for the dissemination of information about an intervention as the ToC map may be a powerful visual tool for describing the key components of an intervention and how it impacts on outcomes. This can be used by other researchers to understand how the intervention worked (for example in systematic reviews [39,40]) and also be used to advocate with policymakers to facilitate the scale-up of successful interventions. Using ToC in this way will be the subject of future research in the projects described in this paper.

As with any approach, there are limitations. The lack of a standardized definition causes confusion and we urge researchers to adopt the definition used by the Aspen Institute [11]. In addition, comprehensive ToC maps may contain a lot of detail with many smaller process preconditions required to achieve impact. Using a detailed ToC with many preconditions and indicators to measure whether that precondition has been achieved may result in an exhaustive list of indicators to measure and a subsequently complex and expensive evaluation plan. This was the case in PRIME, where the demands of conducting a complex evaluation across five countries had to be balanced against the resources required to carry out such an evaluation. As a result, we had to refine the cross-country ToC map to ensure that it contained only the key preconditions and long-term outcomes necessary for the impact to be achieved, and that we only evaluated the key steps in the pathway.

Many of the benefits of the ToC approach derive from the participatory nature of the development of the ToC. If stakeholders are not sufficiently consulted or engaged in the development of the ToC, it is likely that using a ToC becomes yet another box to tick rather than a deeper exploration of the pathways to achieve impact [13]. This may particularly be the case where the decision to develop a ToC is made by the funder rather than seen as an integral part of program development, as shown by the use of ToC as part of the evaluation of the Health Action Zones in the UK. However, more than three quarters of the initiatives did not develop a ToC map as implementers felt that the development of a ToC was taking resources away from implementation [18]. In our experience, having a nominated ToC champion on the research team who is tasked with overseeing the ToC process and driving it forward throughout the project, is critical to the success of the approach.

Our experiences resonate with other examples of applications of theory-driven evaluation approaches, including

ToC, which are reported in the literature. Afifi *et al.* [41] found that using a participatory approach to developing a logic model as the basis for a mental health promotion intervention for youth in a refugee community in Beirut improved the design of their intervention. In their program, a community youth committee was involved in the development of the logic model and provided input into the content and delivery format of the intervention resulting in a more relevant, feasible and sustainable intervention. Similarly, Hernandez and Hodges [42] used ToC developed with stakeholders to organize services for youth in contact with the juvenile justice system. They found that ToC assisted with creating a shared vision among stakeholders which promoted service integration across a variety of sectors. This also allowed planners to envisage what is expected within a community and how the actions of stakeholders can bring this about. Other experiences also highlight that ToC can assist with structuring and prioritizing the evaluation of complex interventions [17,43-46]. However, few provide enough detail to understand how ToC informed both the design of the program and the subsequent evaluation.

It is still in the early stages. While we have tested the use of ToC in three research projects across six countries, these are all mental health programs in low- and middle-income countries, and none have completed the evaluation, analysis or dissemination of the evaluation. Further research is needed in other settings, for other types of complex interventions, and into the usefulness of ToC as a framework for analysis and dissemination of results.

Conclusions

This paper is the first to describe the use of ToC in conjunction with the MRC framework for the development and evaluation of complex interventions, and to provide three case studies testing this approach. Indications from our initial experiences are that, used in conjunction with the MRC framework, ToC may be a useful tool to improve the development and evaluation design of complex interventions in research projects. We urge researchers to consider using this approach and to evaluate its usefulness within a research context.

Endnotes

[a]http://www.theoryofchange.org/

[b]http://www.centreforglobalmentalhealth.org/projects-research/share-south-asian-hub-advocacy-research-and-education-mental-health

[c]See for example: Patient and Public Involvement http://www.ccf.nihr.ac.uk/PPI/Pages/default.aspx and the James Lind Alliance http://www.lindalliance.org/.

Additional files

Additional file 1: Summary Theory of Change from the RISE trial.

Additional file 2: Summary Theory of Change from the PRogramme for Improving Mental health carE (PRIME).

Abbreviations

ToC: Theory of Change; MRC: Medical Research Council; DfID: Department for International Development.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

MDS conceived the idea for this project and developed the initial use of ToC for the design and evaluation of complex interventions in conjunction with the MRC framework, with support from LL, EB, CL and VP. EB, MDS, CL and VP developed the use of ToC in the PRIME program. NC, VP and MDS developed the use of ToC in the SHARE trial. LA and MDS developed the use of ToC in the RISE trial. MDS wrote the first draft of the paper. LA wrote Case Study 1 and EB Case Study 2. All authors revised and gave final approval to the paper.

Acknowledgements

The concept was developed as part of a wider research consortium, the PRogramme for Improving Mental health carE (PRIME), funded by the UK Department for International Development (DfID) for the benefit of LMIC (HRPC10). It was further developed as part of a project to develop a Theory of Change for Grand Challenges Canada Global Mental Health program. The South Asian Hub for Advocacy, Research and Education on mental health program is funded by the US National Institutes of Mental Health (NIMH) (grant number: 1U19MH095687-01). MDS is funded by an LSHTM/Wellcome Trust Fellowship and Grand Challenges Canada, EB by the UK Department for International Development (HRPC10), LA is supported by the Wellcome Trust (grant number: 100142/Z/12/Z), CL by the UK Department for International Development (HRPC10) and NIMH (grant number: U19MH095699), LL by Grand Challenges Canada, and VP by a Wellcome Trust Senior Research Fellowship in Tropical Medicine. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The content is solely the responsibility of the authors and does not necessarily represent the official views of any of the funders.

We are grateful to all of our project partners and collaborators on the SHARE, PRIME, RISE and Grand Challenges Canada projects for working with us to develop the use of Theory of Change for the design and evaluation of complex interventions.

Author details

[1]Centre for Global Mental Health, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK. [2]Department of Psychiatry and Mental Health, Alan J Flisher Centre for Public Mental Health, University of Cape Town, 46 Sawkins Road, Rondebosch, 7700 Cape Town, South Africa. [3]Sangath, Alto-Porvorim, Bardez, Goa 403521, India.

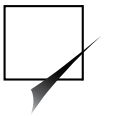
Received: 29 January 2014 Accepted: 16 June 2014
Published: 5 July 2014

References

1. Craig P, Dieppe P, Macintyre S, Nazareth I, Petticrew M: **Developing and evaluating complex interventions: the new Medical Research Council guidance.** *Br Med J* 2008, **337:**a1655.
2. Craig P, Petticrew M: **Developing and evaluating complex interventions: reflections on the 2008 MRC guidance.** *Int J Nurs Stud* 2013, **50:**585–587.
3. Anderson R: **New MRC guidance on evaluating complex interventions.** *Br Med J* 2008, **337:**a1937.
4. Bonell C, Fletcher A, Morton M, Lorenc T, Moore L: **Realist randomized controlled trials: a new approach to evaluating complex public health interventions.** *Soc Sci Med* 2012, **75:**2299–2306.
5. Ling T: **Evaluating complex and unfolding interventions in real time.** *Evaluation* 2012, **18:**79–91.

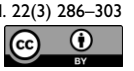
6. Coryn CLS, Noakes LA, Westine CD, Schrote DC: **A systematic review of theory-driven evaluation practice from 1990 to 2009.** *Am J Eval* 2011, **32:**199–226.
7. Kirkpatrick DL: **Techniques for Evaluating Training Programmes".** *J Am Soc for Train Dev* 1959, **11:**1–13.
8. Chen HT: *Theory Driven Evaluation.* Thousand Oaks, California: Sage; 1990.
9. James C: *Theory of Change Review: A report comissioned by Comic Relief*; 2011.
10. Weiss CH: **Nothing As Practical As Good Theory: Exploring Theory-Based Evaluation For Comprehensive Community Initiatives For Children And Families.** In *New Approaches to Evaluating Community Initiatives Volume 1 Concepts, Methods and Contexts.* Edited by Connell JP, Kubisch AC, Schorr LB, Weiss CH. Washington DC: The Aspen Institute; 1995:65–92.
11. Connell JP, Kubisch AC: **Applying a Theory of Change Approach to the Evaluation of Comprehensive Community Initiatives: Progress, Prospects, and Problems**. In *New Approaches to Evaluating Community Initiatives, Volume 2: Theory, Measurement, and Analysis.* Washington DC: The Aspen Institute; 1998.
12. Anderson A: *The Community Builder's Approach to Theory of Change. A Practical Guide to Theory Development.* New York: The Aspen Institute; 2005.
13. Vogel I: *Review of the use of Theory of Change in International Development.* London: UK Department of International Development; 2012.
14. Sullivan H, Barnes HM, Matka E: **Building collaborative capacity through `Theories of Change': early lessons from the evaluation of Health Action Zones in England.** *Evaluation* 2002, **8:**205–226.
15. Mason P, Barnes M: **Constructing theories of Change: methods and sources.** *Evaluation* 2007, **13:**151–170.
16. Weitzman BC, Mijanovich T, Silver D, Brecher C: **Finding the impact in a messy intervention: using an integrated design to evaluate a comprehensive citywide health initiative.** *Am J Eval* 2009, **30:**495–514.
17. Weitzman BC, Silver D, Dillman KN: **Integrating a comparison group design into a Theory of Change evaluation: the case of the urban health initiative.** *Am J Eval* 2002, **23:**371–385.
18. Cole M: **The Health Action Zone initiative: lessons from Plymouth.** *Local Govern Stud* 2003, **29:**99–117.
19. Kernik D: **Wanted—new methodologies for health service research. Is complexity theory the answer?** *Fam Pract* 2006, **23:**385–390.
20. Ajzen I: **The theory of planned behavior.** *Organ Behav Hum Decis Process* 1991, **50:**179–211.
21. Murray M, Treweek S, Pope C, MacFarlane A, Ballini L, Dowrick C, Finch T, Kennedy A, Mair F, O'Donnell C, Ong B, Rapley T, Rogers A, May C: **Normalisation process theory: a framework for developing, evaluating and implementing complex interventions.** *BMC Med* 2010, **8:**63.
22. Department for International Development: *How to Note: Guidance on using revised Logical Framework.* London; 2011.
23. Clark H, Andersen A: *Theories of Change and Logic Models: Telling Them Apart. 2004 Presented at the American Evaluation Association Conference.* Atlanta, Georgia: 2004.
24. Rahman A, Malik A, Sikander S, Roberts C, Creed F: **Cognitive behaviour therapy-based intervention by community health workers for mothers with depression and their infants in rural Pakistan: a cluster-randomized controlled trial.** *Lancet* 2008, **372:**902–909.
25. Taplin D, Rasic M: *Theory of Change Technical Papers: A series of papers to support development of theories of change based on practice in the field.* New York: ActKnowledge; 2013.
26. Taplin DH, Rasic M: *Facilitator's Source Book: Source Book for facilitators leading Theory of Change development sessions.* New York: ActKnowledge; 2012.
27. Grant A, Treweek S, Dreischulte T, Foy R, Guthrie B: **Process evaluations for cluster-randomized trials of complex interventions: a proposed framework for design and reporting.** *Trials* 2013, **14:**15.
28. Bagozzi R, Yi Y: **Specification, evaluation, and interpretation of structural equation models.** *J Acad Market Sci* 2012, **40:**8–34.
29. Robinson S: *Simulation - The practice of model development and use.* Chichester, UK: Wiley; 2004.
30. Bonabeau E: **Agent-based modeling: methods and techniques for simulating human systems.** *Proc National Acad Sci* 2002, **99:**7280–7287.
31. Sterman J: **System dynamics modeling: tools for learning in a complex world.** *Calif Manage Rev* 2001, **43:**8–25.
32. Campbell NC, Murray E, Darbyshire J, Emery J, Farmer A, Griffiths F, Guthrie B, Lester H, Wilson P, AL K: **Designing and evaluating complex interventions to improve health care.** *Br Med J* 2007, **334:**455–459.
33. WHO: *Community based rehabilitation: CBR guidelines.* Geneva: World Health Organization; 2010.
34. Chatterjee S, Naik SS, John S, Dabholkar H, Balaji M, Koschorke M, Varghese M, Thara R, Weiss HA, Williams P, McCrone P, Patel V, Thornicroft G: **Effectiveness of a community-based intervention for people with schizophrenia and their caregivers in India (COPSI): a randomized controlled trial.** *Lancet* 2014, **383:**1385–1394.
35. Lund C, Jordans M, Petersen I, Bhana A, Kigozi F, Prince M, Thornicroft G, Hanlon C, Kakuma R, McDaid D, Saxena S, Chisholm D, Raja S, Kippen-Wood S, Honikman S, Fairall L, Patel V: **PRIME: A programme to reduce the treatment gap for mental disorders in five low- and middle-income countries.** *PLoS Med* 2012, **9:**e1001359.
36. Breuer E, De Silva M, Fekadu A, Luitel N, Murhar V, Nakku J, Petersen I, Lund C: **Using workshops to develop Theories of Change in five low and middle income countries: lessons from the Programme for Improving Mental Health Care (PRIME).** *Intl J Ment Health Syst* 2014, **8:**15.
37. Hanlon C, Luitel NP, Kathree T, Murhar V, Shrivasta S, Medhin G, Ssebunnya J, Fekadu A, Shidhaye R, Petersen I, Jordans M, Kigozi F, Thornicroft G, Patel V, Tomlinson M, Lund C, Breuer E, De Silva M, Prince M: **Challenges and opportunities for implementing integrated mental health care: a district level situation analysis from five low- and middle-income countries.** *PLoS One* 2014, **9:**e88437.
38. Hawe P, Shiell A, Riley T, Gold L: **Methods for exploring implementation variation and local context within a cluster randomized community intervention trial.** *J Epidemiol Community Health* 2004, **58:**788–793.
39. Segal L, Sara Opie R, Dalziel K: **Theory! The missing link in understanding the performance of neonate/infant home-visiting programs to prevent child maltreatment: a systematic review.** *Milbank Q* 2012, **90:**47–106.
40. Anderson L, Anderson L, Petticrew M, Rehfuess E, Armstrong R, Ueffing E, Baker P, Francis D, Tugwell P: **Using logic models to capture complexity in systematic reviews.** *Res Synth Meth* 2011, **2:**33–42.
41. Afifi R, Makhoul J, El Hajj T, Nakkash R: **Developing a logic model for youth mental health: participatory research with a refugee community in Beirut.** *Health Policy Plan* 2011, **26:**508–517.
42. Hernandez M, Hodges S: **Applying a theory of change approach to interagency planning in child mental health.** *Am J Community Psychol* 2006, **38:**165–173.
43. Chandani Y, Noel M, Pomeroy A, Andersson S, Pahl M, Williams T: **Factors affecting availability of essential medicines among community health workers in Ethiopia, Malawi, and Rwanda: solving the last mile puzzle.** *Am J Trop Med Hyg* 2012, **87:**120–126.
44. Hawkins J, Brown E, Oesterle S, Arthur M, Abbott R, Catalano R: **Early effects of Communities That Care on targeted risks and initiation of delinquent behavior and substance use.** *J Adolesc Health* 2008, **43:**15–22.
45. Bickman L: **The application of program theory to the evaluation of a managed mental health care system.** *Eval Program Plann* 1996, **19:**111–119.
46. Van Belle S, Marchal B, Dubourg D, Kegels G: **How to develop a theory-driven evaluation design? Lessons learned from an adolescent sexual and reproductive health programme in West Africa.** *BMC Public Health* 2010, **10:**741.

doi:10.1186/1745-6215-15-267
Cite this article as: De Silva *et al.*: Theory of Change: a theory-driven approach to enhance the Medical Research Council's framework for complex interventions. *Trials* 2014 **15**:267.

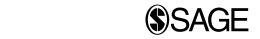


Article

Evaluation
2016, Vol. 22(3) 286–303



DOI: 10.1177/1356389016652743
evi.sagepub.com



Realist complex intervention science: Applying realist principles across all phases of the Medical Research Council framework for developing and evaluating complex interventions

Adam Fletcher
Cardiff University, UK

Farah Jamal
UCL Institute of Education, UK

Graham Moore
Cardiff University, UK

Rhiannon E. Evans
Cardiff University, UK

Simon Murphy
Cardiff University, UK

Chris Bonell
London School of Hygiene & Tropical Medicine, UK

Abstract
The integration of realist evaluation principles within randomised controlled trials ('realist RCTs') enables evaluations of complex interventions to answer questions about *what works, for whom and*

Corresponding author:
Adam Fletcher, School of Social Sciences, Cardiff University, 1–3 Museum Place, Cardiff CF10 3BD, UK.
Email: FletcherA@cf.ac.uk

under what circumstances. This allows evaluators to better develop and refine mid-level programme theories. However, this is only one phase in the process of developing and evaluating complex interventions. We describe and exemplify how social scientists can integrate realist principles across *all phases* of the Medical Research Council framework. Intervention development, modelling, and feasibility and pilot studies need to theorise the contextual conditions necessary for intervention mechanisms to be activated. Where interventions are scaled up and translated into routine practice, realist principles also have much to offer in facilitating knowledge about longer-term sustainability, benefits and harms. Integrating a realist approach across all phases of complex intervention science is vital for considering the feasibility and likely effects of interventions for different localities and population subgroups.

Keywords
Complex interventions, complex systems, realism, evaluation, pilot trials, randomised controlled trials, public health

Introduction

The original UK Medical Research Council (MRC) framework for evaluating complex interventions recommended sequential phases of development, feasibility testing and evaluation, culminating in the estimation of an effect size via a randomised controlled trial (RCT), prior to wider implementation (Campbell et al., 2000). This emphasis on aggregate effectiveness, reflected within many subsequent trials of complex public health interventions, has left trialists open to critiques from 'realist evaluators' (for example, Pawson, 2013) that trials oversimplify causality, and are fundamentally unsuited to the evaluation of complex interventions. Effect sizes may tell us that an intervention helped more people than it harmed in the time and place it was delivered, but often tell policymakers and practitioners little regarding how findings might be applied in new settings or to other populations (Cartwright and Hardie, 2012). An emphasis purely on aggregate effectiveness also means that we risk developing, evaluating and recommending interventions for implementation that have small population-level benefits at the expense of widening existing inequalities (Whitehead 2007).

However, the fact that trialists have not historically considered these issues sufficiently does not mean that they cannot. While often presented as opposing factions (Marchal et al., 2013; Pawson and Tilley 1997), experimental social science is highly compatible with the methodological principles and epistemological assumptions of critical realism which underpin realist evaluation (Bonell et al., 2012, 2013a). Critical realism is a philosophy of science founded on the stratification of social reality into the domains of the real, the actual and the observable (Pawson, 2013). Critical realism seeks to support social scientific investigation through a recognition that the object of such investigation must have real, internal mechanisms that can be *actualised* to produce particular social outcomes (Bhaskar, 2008). Evaluation, including through experimental designs, directly supports the scientific observation of such mechanisms, which are activated in certain contexts of the actual, to explain patterns of social causation and problems in the domain of the real (Bonell et al., 2013a).

Realist evaluation focuses on building, testing and refining middle-range theories regarding complex casual mechanisms and how these interact with individuals' agency and social

context to produce outcomes (Hawkins, 2014; Pawson, 2013). The term 'middle-range theory' was developed to distinguish grand social theories (e.g. functionalism) from the process of integrating theory and empirical research to explain patterns of social behaviour and outcomes in a particular social setting (Merton, 1968). The development and testing of theories about context–mechanism–outcome (CMO) configurations within realist evaluation is one such example of middle-range theory and research (Pawson, 2013), and this process can build on programme 'logic models' that define the components and intended mechanisms of action of specific interventions (Bonell et al., 2012). The most recent MRC guidance on evaluating complex interventions, while maintaining that RCTs should be used to test effectiveness where possible, placed increased emphasis on the use of evaluation to build theory and understand causal mechanisms (Craig et al., 2008a), though the role of context in shaping implementation and causal processes is only briefly mentioned. In particular, aspects of this guidance focussed on intervention development pay no attention to context. Unlike with realist evaluation, there is little emphasis on developing and testing theories.

An emergent field of enquiry within evaluation, which is highly compatible with realist principles and foregrounds the role of context in understanding complex interventions, is complex systems science (Hawe et al., 2009; Westhorp, 2012). Indeed, the MRC guidance has been criticised by some for the use of the term 'complex' in the absence of engagement with complexity theories and thinking (Anderson 2008; De Silva et al., 2014). At present, the MRC guidance conceives complexity largely in terms of synergies between intervention components (for example, the added value of combining an educational component with an environmental component). However, Hawe (2015a), who has advocated the use of RCTs in evaluating complex interventions (Hawe et al., 2004; Shiell et al., 2008), argues that we should conceive complexity in terms of how interventions interact with their contexts. A social intervention represents a disruption to complex systems, or attempts to change the dynamics of the systems in which they are delivered, and hence pre-existing contextual factors will shape what is delivered, how it will work, and for whom (Hawe et al., 2009). Using the example of early intervention programmes, Westhorp (2013) has illustrated the compatibility of 'complexity-consistent theory' for refining mid-level programme theories about mechanisms of actions and the contexts that activate them.

Thus, there is an inherent compatibility of complex systems science, critical realism and realist evaluation in their mutual commitment to understanding causality within complex environments. Ontologically, these approaches are consistent that causality should be understood as always dependent on the whole context of an intervention, including the complex and emergent systems within which it is embedded (Byrne, 2013). That is to say, causation is a consequence of multiple factors rather than any single specific factor, and will operate in different ways such that the same outcome may be generated by different causal combinations in different contexts. There is also substantial overlap between a complexity approach to evaluation and realist evaluation, due to their explicit concern with social theory and focus on understanding the interplay of agency and structure (Byrne, 2013).

Progress is being made in integrating complex systems science and realist evaluation principles with RCTs through 'realist RCT' designs, to allow evaluators to go beyond simply asking 'does it work' and towards more nuanced consideration of *what works, for whom and under what circumstances* (Bonell et al., 2012). Large-scale realist RCTs are now being undertaken in the UK (for example, Bonell et al., 2014) and sub-Saharan Africa (for example, Chandler et al., 2013). New MRC guidance on integrating process evaluation within trials of

complex interventions also endorses the use of RCTs that integrate qualitative data collection and analysis focussed on the interactions between mechanisms, context and outcomes (Moore et al., 2014, 2015). However, effectiveness trials are only one phase within the process of developing and evaluating public health interventions. In order for realist RCTs to deliver health improvement benefits via developing well-theorised, effective, scalable health improvement interventions, it is vital that other phases of intervention development and refinement are also as clearly focussed on generating knowledge about their mechanisms of action and how these can interact with social context to produce various outcomes.

Complex intervention science phases

The 2008 update of the MRC guidance for complex intervention development and evaluation provides a four-phase, cyclical framework advising health researchers to answer a range of sequential questions regarding complex intervention theory, feasibility and acceptability, effectiveness and cost-effectiveness, and sustainability (Craig et al., 2008a). The first phase (intervention development) involves the development of an intervention's theoretical rationale, often depicted in a 'logic model' describing inputs that the intervention involves, the processes that these initiate, and the mechanisms via which these are intended to realise positive outcomes. This phase should identify underpinning 'active ingredients' and how intervention components are expected to synergistically interact with one another, and with the context of delivery (although less emphasis is given to this), to generate outcomes (both intended and unintended) (Bonell et al., 2015).

The subsequent feasibility and piloting phase includes testing the feasibility and acceptability of the proposed intervention and its evaluation methods. Although the exact distinction between feasibility and pilot studies is contested (Lancaster, 2015), pilot studies may simply be a smaller version of the main trial, aiming to implement the intervention and its trial on a smaller scale (for example, with smaller samples, in fewer sites and/or for shorter follow-up periods), while feasibility studies may focus only on select intervention or trial elements about which there is particular uncertainty. Further refinements may be made to the intervention theory after this phase to optimise the intervention design, logic model and the proposed evaluation design prior to testing effectiveness and cost-effectiveness.

Once a well-theorised intervention has been developed and feasibility questions addressed, RCTs are recommended to examine their effectiveness (and cost-effectiveness) whenever randomisation is practicable (Craig et al., 2008a). Finally, 'implementation studies' are also needed to address the scale-up of interventions into routine practice (Craig et al., 2008a). The cumulative effect of these processes should be the generation of a strong theoretical and evidence base for public health intervention which provides greater confidence that outcomes observed during trials can be replicated in real-world settings, and which supports the ongoing cycle of developing and evaluation complex interventions.

This article outlines how realist evaluation principles have much to offer public health intervention science, not only for trials of effectiveness but also across all phases of public health intervention science, from intervention development, feasibility and pilot studies to post-evaluation scale-up studies. For example, as the number and range of feasibility and pilot studies proliferates (Arain et al., 2010; Lancaster, 2015), a realist lens can be applied to such studies to address questions regarding not only what is feasible and acceptable in general, but also for whom and under what circumstances, and place much more emphasis on exploring

potential mechanisms of action (i.e. the intermediate processes triggered by the introduction of an intervention, which give rise to intended, and unintended, consequences) and how these may vary by context prior to large-scale realist RCTs. This is vital in ensuring that we are clear via what mechanisms and in what contexts interventions are expected to work, and for whom, and focus later phases of evaluation on interventions that have potential to be deliverable in the most salient settings, effective for key populations, and are scalable. Once realist RCTs of complex interventions have demonstrated their effectiveness, subsequent realist evaluations of their scale-up should enable us to further refine our understanding of how these interventions play out in an even greater diversity of contexts. This will better inform attempts to adapt implementation to local conditions while ensuring consistency with the core theoretical principles of the intervention.

Some of the authors of the revised MRC guidance have subsequently argued that approaches such as complex systems science and realist evaluation may become routine within public health evaluation methods once sufficient empirical examples are available to guide practice (Craig and Petticrew, 2013). This article draws on new case examples of realist studies across the different phases within the latest MRC guidance (Craig et al., 2008b) to provide guidance on the theoretical and methodological process of integrating a realist approach throughout this cycle of intervention development and evaluation. Each phase of intervention science is considered in turn: from intervention development and feasibility and pilot studies, to subsequent evaluations of intervention effectiveness, and implementation studies of scaled-up interventions. We conclude by discussing what structures and partnerships are also required to facilitate realist intervention science, such as the development of specialist social science trials infrastructure to embed these principles within public health evaluation science, and further investment in transdisciplinary research networks to support the quantity, quality and relevance of realist intervention science (Glasgow et al., 2003; Stokols, 2006).

Intervention development and modelling

Within the revised MRC guidance, there is relatively little attention paid to the developmental phase of the complex intervention cycle (Craig et al., 2008a,b). Other frameworks and toolkits have been developed to specifically support intervention development but these tend to ignore the complexity of multi-component, and particularly multi-level, approaches to health improvement and also the importance of considering context (Hawe, 2015b). For example, the literature providing guidance on the development of intervention logic models is still informed by simple, linear behaviour–determinant–intervention (BDI) toolkits (e.g. Kirby, 2004) and ignores how implementation and causal pathways may vary by context (for example, 'intervention mapping' as proposed by Bartholomew et al., 2011).

More recently, theoretically orientated tools have been developed, such as the 'Behaviour Change Wheel' (Michie et al., 2011) and the 'Theory of Change tool' (De Silva et al., 2014) with the aim of improving public health intervention development. However, these focus on helping researchers and practitioners categorise and label intervention inputs and activities more systematically, which overprivileges parsimony and oversimplifies complex social realities. These tools also do not engage with a realist approach focussed on theorising mechanisms nor how these vary by context. These approaches also tend to suggest an idealised and highly linear sequence in which, for example, all objectives and pathways are pre-specified prior to designing components and planning implementation, which, first, ignores the potential of

retrospective theoretical modelling of existing interventions and, second, overlooks the likelihood that all mid-level programme theories will need to be iteratively tested and refined in the light of subsequent pilot and evaluation findings.

Addressing these existing gaps in the literature and via engagement with a realist lens, we recommend further development and use of the following three methods to support intervention development and modelling: mixed-methods evidence synthesis; formative mixed-method, multi-case-study research; and, pragmatic formative process evaluation. These methods would support the development of more three-dimensional (3-D) logic models, which focus not only on complex the pathways from (1) inputs to (2) outcomes but also the (3) contextual dimensions that activate or mitigate causal processes. Intervention logic models (referred to as implementation models by Weiss, 1995) have typically focussed on defining the components and mechanisms of specific interventions within a very particular setting and paid relatively little attention to how mechanisms interact with context and produce potentially contradictory processes and outcomes in different localities and for various populations subgroups (Bonell et al., 2012; Moore et al., 2015). The inclusion of a contextual dimension within the logic models at the intervention development stage would in turn support the subsequent phases of realist evaluation, which are outlined later in this article.

Mixed-method evidence synthesis

The process of designing more theoretically driven interventions and specifying potential CMO configurations has been hindered by the dominant paradigm within evidence syntheses: systematic reviews still typically focus on synthesising only quantitative studies answering questions about 'what works' at the expense of understanding how, in what context and for whom (Pawson, 2013; Petticrew, 2015). These evidence reviews therefore still typically only focus on accrediting public health policies and interventions as 'effective' (or otherwise). Methods such as meta-analysis traditionally aggregate across studies to derive overall effect sizes, rather than exploring how and why trials of similar interventions produce different outcomes in different contexts. The dominance of such reductionist methods is associated with the rise of intervention-comparison websites (similar to price-comparison websites), such as the *Blueprints Youth Programmes* resource developed in the USA (http://www.blueprintsprograms.com/) and the UK *Investing in Children* database (http://investinginchildren.eu/), which accredit lists of 'effective' interventions without consideration of which contexts such interventions might be suitable.

Mixed-methods reviews have similarities with mixed-methods primary research, thus there are many ways in which the products of different syntheses methods can be combined to overcome the limitations with traditional systematic review methods. 'Realist reviews' have been suggested as an alternative (or adjunct) to address the lack of focus on CMO configurations in current evidence syntheses (Pawson et al., 2005). However, although realist review guidelines include a stronger focus on examining context as well as outcomes (Wong et al., 2013) and can provide a conceptual platform prior to complex intervention development (Pearson et al., 2015a), they are more open ended and often not do involve an a priori protocol. Such protocols are necessary to minimise bias and retain practical focus, and this has limited the potential of realist reviews to support the development of practical, theoretically driven, population-level health improvement interventions. As with realist trials (Bonell et al., 2012; Jamal et al., 2015), it is possible for systematic reviews to be guided by a priori protocols while being

mixed method and thus more attentive to mechanism and context. To do this, reviews can continue to synthesise evidence of overall effects from RCTs and quasi-experimental studies (including via meta-analysis where appropriate) while also undertaking other syntheses to better understand how interventions work and how this might vary with context. There are two main ways of doing this.

First, reviews can synthesise information on theories of change and evidence on intervention processes to develop hypotheses about the mechanisms via which interventions are intended to work, as well as how implementation and effectiveness might be affected by the characteristics of different populations and places. For example, two recent mixed-methods reviews – one examining how the school environment and school–environment interventions influence health, and one examining the effects of community-based positive youth development (PYD) interventions – have synthesised intervention theories and the findings from process evaluation reports as well as estimates of intervention effects to hypothesise how school environment and PYD interventions can improve health, for whom and in what contexts (Bonell et al., 2013b, 2016). A realist systematic review and synthesis of studies examining the process of implementing health programmes in schools also highlights the benefits of reviewing process data systematically to develop programme theories and support intervention design (Pearson et al., 2015b). This method allowed the authors to identify transferable mechanisms that support implementation when preparing for, and introducing, new programmes in a school.

Second, reviews can use meta-regression or qualitative comparative analysis (QCA) (Ragin et al., 2006; Thomas et al., 2014) to examine how intervention effects vary according to the characteristics of settings or populations, or examine intervention effects on potential mediators and whether these might account for effects on primary outcomes. With both of the school environment and PYD reviews cited above, the intention was to use the hypotheses derived from syntheses of theories of change and process evidence to inform selection of which moderator and mediator variables to examine in syntheses of outcome evaluations. In neither case was this possible because the included outcome evaluations did not report potential moderators or mediators consistently enough to allow syntheses to examine these. However, other reviews, while not using preliminary syntheses of theoretical literature and process evidence to inform hypotheses, have been able to test what contextual factors appear to moderate intervention effectiveness. For example, a review and meta-analysis of criminal justice interventions by Lipsey (2009) examined how the site of delivery moderated effectiveness. QCA has also been tested and allowed reviewers to go beyond basic, narrative synthesis of integrated process evaluations and identify key intervention characteristics and how effects may occur (for example, Thomas et al., 2014). Such methods of evidence synthesis will be facilitated as more studies adopt a realist lens, as outlined in the discussion.

Formative case studies

As well as mixed-methods systematic reviews to identify the relevant theoretical and evidence base, before new interventions are piloted it is often useful to undertake formative, mixed-method case-study research to understand their socio-ecological context, explore potential intervention delivery and hypothesise mechanisms of action. Such formative case studies can employ purposive sampling to provide contextual diversity, informed by initial theories, and generate insights regarding how these contexts might interact with intervention mechanisms to influence outcomes for different groups.

One example of this design is a current formative study to develop and model a new intervention to be delivered in further education (FE) colleges to promote safe sex and relationships among 16–19-year-olds. Six FE colleges in England and Wales were purposively sampled according to type and size of institution. A phased approach to data collection and analysis supports the consideration of CMO. First, focus groups and interviews have been used to explore the views of students, teachers, managers and sexual health service providers on how interventions deliverable within FE colleges might work to improve relationships and sexual health. Second, informed by these data, a larger cross-section of students and staff were surveyed to develop theories about how these mechanisms might interact with context to play out differently in different settings and/or with different groups of students (for example by gender, sexuality, socioeconomic status (SES) and/or baseline sexual risk). Finally, findings from these elements will be brought together to refine a 3-D intervention logic model which incorporates consideration of CMO configurations.

The design and development of a new film-based intervention targeting teenage men to prevent unintended pregnancy has also involved formative, mixed-methods research in a range of settings (Aventin et al., 2015). To develop a theoretical understanding of the phenomenon of unintended teenage pregnancy in relation to young men – who are not typically targeted by teenage pregnancy prevention interventions – a mix of methods was necessary, including consultations with schools, focus groups and a survey to assess the views of a wider cross-section of young men aged 14–17 about potential intervention components. A strength of this study is that it went beyond the basic MRC guidance on developing complex interventions by also explicitly addressing contextual complexities through engaging a range of the target group (young men) across a range of settings (schools) (Aventin et al., 2015).

Pragmatic process evaluations

The development of new interventions and modelling of theories of change can also be enhanced by pragmatic process evaluations of interventions already in routine practice (Evans et al., 2015a). Although such evaluations remain somewhat rare, these designs allow us to move beyond the theorisation of how a postulated theory of change may play out in real-world settings as intervention mechanisms are already interacting with contextual characteristics across a range of settings: the ‘C’ element of CMO configuration is already privileged within pragmatic, formative evaluations (Evans et al., 2015a).

These evaluations allow for the examination of mechanisms not only of intended benefits but also unanticipated consequences, including unintended harms. For example, a pragmatic formative process evaluation of a school-based social and emotional learning intervention identified a number of iatrogenic effects as a consequence of the stigmatising referral processes and negatively labelling young people (Evans et al., 2014). Through using a mixture of direct observations and interviews with multiple stakeholders to capture their different perspectives, these studies also provide insights into the organisational-level barriers and facilitators of implementation (Evans et al., 2015b). Whereas the MRC progression framework has tended to address implementation and translational issues at the point of scale-up following a trial, pragmatic process evaluation of existing interventions allow this to be theorised and empirically explored from the start, which will help to ensure intervention development studies have external, and socio-ecological, validity and supports more sustainable implementation procedures.

Our suggestion is not that resources should be used to retrospectively theorise all existing interventions on an exhaustive basis. However, once existing interventions are deemed to warrant outcome and process evaluation they should be first subjected to pragmatic formative process evaluation to help develop the intervention logic model, model realist CMO hypotheses and, if necessary, refine delivery methods prior to larger-scale evaluation and scale-up. Without a clear theory of change, subsequent evaluations employing a realist perspective will be of more limited value. One example of where an existing but under-theorised intervention was subjected to pragmatic process evaluation was the Welsh National Exercise Referral Scheme (NERS) (Murphy et al., 2012). Theoretically informed analyses of the trial data were able to examine variations in health benefits across different groups, and contextual interactions, which are described below ('Realist RCTs') as an illustration of the benefits of integrating realist principles across multiple evaluation phases.

Realist feasibility and pilot studies

Feasibility and pilot studies should also apply a realist approach to explore implementation and potential mechanisms of action in a range of contexts prior to larger effectiveness trials. Following the development of MRC guidance on complex interventions (Campbell et al., 2000; Craig et al., 2008a), the volume of feasibility and pilot studies, particularly pilot RCTs, has increased markedly (Arain et al., 2010; Lancaster, 2015). Such preliminary studies of theoretically informed interventions provide an opportunity to examine barriers and facilitators to implementation in a range of settings, to explore the views of those involved, and to refine and optimise the intervention design, logic model and trial methods prior to realist RCTs. However, to date, pilot RCTs have often only answered relatively crude, binary questions about whether a specific complex intervention is feasible and acceptable, or not.

The dominance of such binary assessments is now reflected in the widespread use of binary 'progression criteria', including by funders, to determine whether a subsequent, larger evaluation is warranted (e.g. Newbury-Birch et al., 2014). Feasibility and pilot studies should instead assess what is feasible and acceptable for whom and under what circumstances, aiming to refine hypotheses about potential mechanisms of action and how these might vary by context, and pilot the methods and measures that can capture these. Several realist strategies have been used and should be developed and used more widely at this stage in the cycle of intervention development and evaluation to refine intervention theories and support subsequent, large-scale realist evaluation studies testing programme theories.

First, purposive sampling criteria should be used in pilot RCTs to ensure there is sufficient diversity in aspects of context that have been pre-hypothesised to affect feasibility, acceptability and causal mechanisms. It is essential to assess these in a range of contexts, but this rarely happens in practice. One example is a pilot cluster RCT of whole-school restorative approach to prevent bullying and aggression in secondary schools (Fletcher et al., 2015). This study used a purposive sampling matrix to recruit a theoretically informed diversity of schools that varied according to the SES of their students (high/low free school meal eligibility) and inspectorate rating of school 'effectiveness'. This study also purposively sampled a range of more or less experienced intervention delivery staff. In the case of pilot trials in which individuals, rather than clusters, are the unit of allocation, there is still a need to encompass relevant diversity in intervention sites and individuals. Exploration of contextual variation in feasibility and acceptability at this stage also allows researchers to identify ways in which the

intervention delivery might be adapted to different contexts if necessary (while maintaining consistency with underlying theory).

Second, like subsequent realist RCTs (as outlined in Bonell et al., 2012), feasibility and pilot trials provide the opportunity to collect and analyse rich qualitative data to support the refinement of hypotheses about causal pathways to test in subsequent effectiveness trials. Feasibility and pilot studies also do not aim to estimate intervention effects, so research teams can collect much more data, especially qualitative data, from intervention or control groups without concerns about this biasing outcome measurement, for example via Hawthorne effects. A specific progression criterion from pilot to large-scale trials should focus on the refinement of hypotheses in this way.

Third, where appropriate, multi-arm pilot RCTs can be employed to help assess the feasibility, acceptability and potential mechanisms of multiple different interventions, or to pilot multiple intervention components separately. A four-arm cluster randomised pilot trial in 12 secondary schools in south Wales is being used to assess the feasibility, acceptability and potential impacts of different peer-led drug-prevention intervention methods (White et al., 2014). As well as piloting the use of a control group, there are three different 'intervention arms': 'ASSIST', an existing peer-led smoking-prevention intervention targeting year 8 students (aged 12–13); 'ASSIST+Frank', which combines ASSIST with a new informal peer-led drug-prevention adjunct targeting year 9 students (aged 13–14); and 'Frank friends', which is a new stand-alone, informal drug-prevention intervention delivered in year 9. The embedded process evaluation will explore the views of students and school staff regarding the two different pilot methods of delivering peer-led drugs education ('ASSIST+Frank'; 'Frank friends'), and assess implementation fidelity by arm. Depending on the results of piloting, these multi-arm designs may or may not be taken forward as multi-arm, realist RCTs, or it may be decided to merge or remove arms.

Realist RCTs

The term 'realist RCT' has been used to describe large-scale mixed-method trials that combine the advantages of the minimisation of bias in estimating intervention effects via randomisation to a control group, with the ability to theorise the mechanisms underlying these effects as well as how effects differ by social group and place (Bonell et al., 2012, 2013a). This combination means that realist trials maximise internal validity in estimating effects within the trial (and how these are moderated by contextual factors) as well as maximising external validity by developing evidence-based theories about the factors which will promote or limit the effectiveness of the intervention in other settings and with other populations. New MRC process evaluation guidance supports the combination of RCT methods with detailed process evaluation to understand mechanisms and context (Moore et al., 2014, 2015), although there are few examples of such studies to date.

One such example is the Welsh NERS policy trial that built on a pragmatic, formative mixed-method process evaluation to develop the intervention logic model (Moore et al., 2012). In the trial of the NERS, quantitative and qualitative data were then used to test and refine the programme theory. For example, a key hypothesised mechanism for improving physical activity was increased autonomous motivation. Several components targeting this mechanism were not well delivered (Moore et al., 2013). Nevertheless, mediation analyses showed that change in physical activity appeared to be explained by change in autonomous

motivation (Littlecott et al., 2014). It appears from qualitative data that this mechanism was triggered largely by emergent social aspects of the scheme rather than by motivational counselling techniques (Moore et al., 2013). Moderation analyses were also able to examine how effects varied according to subgroups, which found that the programme did not increase physical activity for those patients referred for mental health reasons but did for those referred on the basis of coronary heart disease risk (Murphy et al., 2012). Aforementioned qualitative process data enabled researchers to understand the social processes through which patterning in responses to the intervention emerged.

A realist RCT of a whole-school restorative approach to preventing bullying, which followed the earlier realist pilot RCT described above, is developing and using a three-stage theoretical and methodological process of building and testing mid-level theories (Jamal et al., 2015). First, informed by the findings of the prior pilot study and sociological theory, researchers elaborated the theory of change and specific a priori hypotheses about CMO configurations. Second, emerging findings from the integral process evaluation within the RCT are being used to refine, and add to, these a priori hypotheses before the collection of quantitative, follow-up data. Third, hypotheses are tested using a combination of process and outcome data with quantitative analyses of effect mediation (examining mechanisms) and moderation (examining contextual contingencies). The main output of the RCT is to assess whether the intervention is effective or not, but importantly to also refine and further develop an empirically informed theory of change. This process also supports evaluators to identity both intended and unintended consequences of complex interventions, including through iteratively developing and testing 'dark logic models' (Bonell et al., 2015).

A realist approach to trial design also helps draw greater attention to how aspects of usual care (i.e. the control group condition) may foster mechanisms similar to the intervention in some contexts, which is rarely considered by trialists at present. For example, a meta-analysis of studies examining adherence to HIV care concluded that between-study variation in intervention effectiveness could be explained as much by differences in behaviour change elements in the usual care arms of the included studies as by variation in interventions (De Bruin et al., 2010). More fully theorising comparison-group contexts, as well as building and testing programme theories, is particularly important for fostering appropriate cross-national and cross-cultural replication of programmes. For example, the Family Nurse Partnership programme, an intensive model of prenatal and early childhood home visiting for vulnerable first-time mothers and their children found to be effective in the USA (Olds, 2016), has been replicated and trialled at scale in England with no benefits observed (Robling et al., 2016). Post-hoc theorisation of the programme has focussed on variations in pre-existing community contexts (i.e. control group care), as well as the programme itself, and how the null effects observed in a UK context could be attributed to all mothers having free access to a range of supportive health and social services (Olds, 2016; Robling et al., 2016). To put this another way, the powerful effects observed in the USA appear to be fired through the programme mechanisms interacting with the more 'Darwinian' nature of usual care in that context, with little state support for poor, young mothers for whom the greatest effects were observed.

Scale-up evaluations

Realist approaches can also be applied where interventions are scaled up after successful trials. Evaluations of scale-ups can examine long-term benefits and harms and how these vary

by context. These studies can occur over a wider range of settings, populations and time periods and so have particular strengths in understanding how context shapes outcomes.

One example of this is the evaluation of the scale-up of the Intervention with Microfinance for AIDS and Gender Equity (IMAGE), which did not explicitly use a realist approach but nonetheless embodied some of its key principles. The IMAGE intervention combined group-based lending with gender and HIV education, and facilitated community mobilisation campaigns, targeting women living in poverty in rural South Africa. Following a cluster RCT trial that suggested that this was effective in reducing rates of intimate partner violence (Pronyk et al., 2006), this intervention was scaled up to other rural sites within South Africa. The follow-on scale-up evaluation did not aim to examine effectiveness but built on the process evaluation embedded within the cluster RCT to examine longer-term implementation processes and potential mechanisms in contrasting sites (Hargreaves et al., 2010). This study suggested that community mobilisation components were often not sustainable, particularly in those contexts where women were targeted on the basis of poverty and were socially marginal within the villages in which they lived. Community mobilisation was intended to reduce sexual risk behaviours among women's household members and villagers via a mechanism involving increased critical consciousness of the social determinants of risk. The evaluation's finding that this mechanism may not have been functioning in some contexts provided insights into why IMAGE may only have been effective for the women themselves and enabled refinement of the theory of change.

There are few, if any, other examples of such MRC 'implementation' studies using realist approaches, although there are examples of 'natural experiments' of large-scale interventions using realist approaches (e.g. Humphreys and Eisner, 2014). However, if realist principles come to be applied throughout earlier phases of intervention development and evaluation, there will be greater scope for them to inform wider scale-up and ongoing monitoring.

Discussion

Public health evaluators have typically under-theorised and under-researched how interventions are intended to engage with their social contexts to enact change (Hawe, 2015a; Macintyre and Petticrew, 2000; Moore et al., 2015). If evaluators continue to under-theorise interventions, focus on binary notions of feasibility and acceptability to the neglect of how this is affected by context, and conceptualise complexity only in terms of the number and interaction of intervention components, it is unlikely that their work will amount to a body of intervention theory and scientific knowledge that is useful to policymakers and practitioners who need to know what interventions should be delivered where, how and to whom. A history of what has worked in one time and place cannot be naively treated as a guarantee of future success elsewhere.

While realist RCTs are becoming more common, large-scale outcome evaluations are only one phase in the process of identifying effective, sustainable interventions to improve health. It is also much more difficult to undertake realist RCTs and scale-up studies without earlier phases of development and piloting that develop and refine programme theories and CMO hypotheses. To facilitate a step-change in the quantity and quality of realist RCTs, the development of complex interventions and their theories of change, and preliminary feasibility and pilot studies, should also now adopt a realist focus on context and mechanisms of actions. Purposive sampling is particularly important to ensure a range of contexts are studied at an

early stage and the role of context is therefore theorised alongside the intervention logic model. It is then possible to test hypothesised mechanisms of actions (mediation analyses) and examine how outcomes vary by subgroup and place (moderation analyses) within large-scale realist RCTs, as well refining and building new hypotheses within these trials via qualitative data. In some cases, it may also be possible to test moderated mediation (i.e. whether there is an effect mediated by certain mechanisms only under specific contexts), which remains rare in RCTs.

Adopting such a realist approach across all phases of intervention science is vital for considering the likely effects of interventions on different social groups and addressing inequalities in health and other outcomes. For example, at the stage of developing interventions and modelling their mechanisms, it is important to theorise the processes and outcomes for different sub-populations. If more complex logic models are not developed to embrace system-focussed theory it is unlikely that new interventions will respond effectively to the most entrenched social problems and reduce inequalities (Hawe, 2015b). Feasibility and pilot studies should also include a strong focus on implementation and its acceptability among the most deprived communities to ensure that interventions are feasible and sustainable in such contexts. Realist trials that include moderation analysis to assess variation by SES and place can also help to ensure that we do not develop, evaluate and implement interventions that will exacerbate health inequalities in the future.

The major barrier to formally testing CMO configurations within individual studies are the small sample sizes that trials often use, powered to examine effects on primary outcomes but not necessarily sufficiently powered to detect differences in all secondary or intermediate (process) outcomes. Trials are rarely designed with secondary analyses according to mediators or population subgroup in mind (Petticrew et al., 2012), and clinical trials units often reject such secondary data analyses for fear of false positive results and accusations of 'data dredging' (Davey Smith and Ebrahim, 2002). We would argue strongly that secondary analyses such as those proposed above are important for a full understanding of how interventions work and for whom, although all analyses should be guided by a priori hypotheses set out in protocols. Even where single studies lack the power for such analyses, reporting their results is useful because it then allows these to be used within systematic reviews and meta-analyses. To facilitate this, studies on related interventions and outcomes should as far as possible use common, validated measures.

If RCTs that adopt realist principles become increasingly common, there is also a need for infrastructure investment to develop the procedures for conducting realist analyses (while avoiding data dredging), facilitate and coordinate new studies, and to develop guidance for developing and reporting robust intervention theory of change. First, there is potential for social science trials teams with expertise in realist methodologies to operate within existing clinical trials units to combine expertise in trial statistics and realist approaches for social interventions.

Second, further investment in transdisciplinary research networks – which involve researchers from multiple disciplines, policymakers, practitioners and the public – is required to increase the quantity, quality and relevance of realist intervention science. This transdisciplinary approach limits the problems created by the separation of the research community from policy and practice, including the concentration of academics on efficacy trials that have little impact on practice (Glasgow et al., 2003; Stokols, 2006). Informed by primary care research networks, which facilitated research capacity (Griffiths, et al., 2000) and fostered a culture of practitioner-led enquiry (Thomas and White, 2001), the Public Health Improvement Research

Network (PHIRN) in Wales is one example of a transdisciplinary network that has addressed the limited research capacity, skills and experience of policymakers and practitioners in pragmatic realist complex intervention science. Between 2006 and 2014 PHIRN supported 122 multidisciplinary and multi-sectoral research development groups and secured 72 externally funded research projects focussed on developing and evaluating complex health improvement interventions, including several of the studies cited above (Evans et al., 2014, 2015b; Moore et al., 2013; Murphy et al., 2012; White et al., 2014). As well as increasing the numbers of trials, such co-production can also facilitate mixed-methods reviews of complex interventions (Pearson et al., 2015b; Petticrew et al., 2013) and pragmatic formative studies (Aventin et al., 2015; Evans et al., 2015b). However, there is concern that new UK anti-lobbying regulations may limit, rather than facilitate, knowledge exchange between policymakers and researchers in the future (Smith et al., 2016).

Third, protocol and reporting guidelines should aim to facilitate a step-change towards the realist complex intervention science methods recommended above. For example, trial protocols should include pre-specified moderator and mediator analysis but also allow for iteration in order to refine hypotheses during a trial in light of emerging qualitative data (Bonell et al., 2014; Jamal et al., 2015). Guidance on reporting trials should also include pre-hypothesised mechanism and moderators, for example, within the extension of the CONSORT statement for social and psychological interventions (Mayo-Wilson et al., 2013). Consistent reporting would further support replication studies and systematic reviewers aiming to integrate theory and process data alongside outcome data. Systematic reviewers synthesising social interventions may also value extensions of quality assessment tools (e.g. AMSTAR) that consider key aspects of realist trials principles (e.g. elaborated theory of change, quantitative syntheses of moderator and mediator analyses, and/or QCA). The Cochrane Collaboration's tool for assessing risk and bias should also be reviewed (Higgins et al., 2011); it currently focuses on internal validity with little consideration for how to reliably synthesis evidence about intervention theory and generalisability beyond the trial setting.

These investments in a realist complex invention science infrastructure and new reporting guidelines would support the cost-effective use of evaluation research funding, and the development of policy-relevant evidence to improve health. Significantly, such an approach offers a way to fully theorise and promote progression through the phases in the MRC framework for the development and evaluation of complex interventions. In turn, greater use of realist RCTs and scale-up studies will, in the long term, support new evidence syntheses that answer a wider range of questions about what works, for whom and under what circumstances, and what carries on working once scaled up and sustained. Those developing interventions or describing their intended mechanisms of action can then draw on such reviews to think more clearly about intended mechanisms and how these interact with context to enable outcomes to manifest.

Declaration of conflicting interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The work was undertaken with the support of The Centre for the Development and Evaluation of Complex Interventions for Public Health Improvement (DECIPHer), a UK Clinical Research Collaboration (UKCRC) Public Health Research Centre of Excellence. Joint funding (grant number

MR/KO232331/1) from the British Heart Foundation, Cancer Research UK, Economic and Social Research Council, Medical Research Council, the Welsh Government and the Wellcome Trust, under the auspices of the UKCRC, is gratefully acknowledged.

References

Anderson R (2008) New MRC guidance on evaluating complex interventions. *BMJ* 337(7676): 944–945.

Arain M, Campbell MJ, Cooper CL, et al. (2010) What is a pilot or feasibility study? A review of current practice and editorial policy. *BMC Medical Research Methodology* 10: 67.

Aventin A, Lohan M, O'Halloran P, et al. (2015) Design and development of a film-based intervention about teenage men and unintended pregnancy: Applying the Medical Research Council framework in practice. *Evaluation and Program Planning* 49: 19–30.

Bartholomew LK, Parcel GS, Kok G, et al. (2011) *Planning Health Promotion Programs: An Intervention Mapping Approach*. 3rd ed. San Francisco, CA: Jossey-Bass.

Bhaskar RA (2008) *A Realist Theory of Science*. London: Verso.

Bonell C, Allen L, Christie D, et al. (2014) Initiating change locally in bullying and aggression through the school environment (INCLUSIVE): Study protocol for cluster randomised controlled trial. *Trials* 15: 381.

Bonell C, Fletcher A, Morton M, et al. (2012) Realist randomised controlled trials: A new approach to evaluating complex public-health interventions. *Social Science & Medicine* 75(12): 2299–2306.

Bonell C, Fletcher A, Morton M, et al. (2013a) Methods don't make assumptions, researchers do: A response to Marchal et al. *Social Science & Medicine* 94(1): 81–82.

Bonell C, Hinds K, Dickson K, et al. (2016) What is Positive Youth Development and how might it reduce substance use and violence? A systematic review and synthesis of theoretical literature. *BMC Public Health* 16(1): 135.

Bonell C, Jamal F, Harden A, et al. (2013b) Systematic review of the effects of schools and school environment interventions on health: evidence mapping and synthesis. *Public Health Research* 1(1).

Bonell C, Jamal F, Melendez-Torress GJ, et al. (2015) 'Dark logic': theorising the harmful consequences of public health interventions. *Journal of Epidemiology and Community Health* 69(1): 95–98.

Byrne D (2013) Evaluating complex social interventions in a complex world. *Evaluation* 19(3): 217–228.

Campbell M, Fitzpatrick R, Haines A, et al. (2000) Framework for design and evaluation of complex interventions to improve health. *BMJ* 321: 694–696.

Cartwright N and Hardie J (2012) *Evidence-Based Policy: A Practical Guide to Doing It Better*. New York: Open University Press.

Chandler C, DiLiberto D, Taaka NS, et al. (2013) The PROCESS study: A protocol to evaluate the implementation, mechanisms of effect and context of an intervention to enhance public health centres in Tororo, Uganda. *Implementation Science* 8: 113.

Craig P and Petticrew M (2013) Developing and evaluating complex interventions: Reflections on the 2008 guidance. *International Journal of Nursing Studies* 50(5): 585–587.

Craig P, Dieppe P, Macintyre S, et al. (2008a) Developing and evaluating complex interventions: The new Medical Research Council guidance. *BMJ* 337: a1655.

Craig P, Dieppe P, Macintyre S, et al. (2008b) *Developing and Evaluating Complex Interventions: New Guidance*. London: MRC.

Davey Smith G and Ebrahim S (2002) Data dredging, bias, or confounding. *BMJ* 325(7378): 1437–1438.

De Bruin M, Viechtbauer W, Schaalma HP, et al. (2010) Standard care impact on effects of highly active antiretroviral therapy adherence interventions: A meta-analysis of randomized controlled trials. *Archives of Internal Medicine* 170(3): 240–250.

De Silva M, Breuer E, Lee L, et al. (2014) Theory of Change: A theory-driven approach to enhance the Medical Research Council's framework for complex interventions. *Trials* 15: 267.

Evans R, Murphy S and Scourfield J (2015b) Implementation of a school-based social and emotional learning intervention: understanding diffusion processes within complex systems. *Prevention Science* 16(5): 754–764.

Evans R, Scourfield J and Murphy S (2014) The unintended consequences of targeted interventions: Young people's lived experiences of targeted social and emotional learning interventions. *British Educational Research Journal* 41(3): 381–397.

Evans R, Scourfield J and Murphy S (2015a) Pragmatic, formative process evaluations of complex interventions and why we need more of them. *Journal of Epidemiology and Community Health* 69(10): 925–926.

Fletcher A, Fitzgerald-Yau N, Wiggins M, et al. (2015) Involving young people in changing their school environment to make it safer: Findings from a process evaluation in English secondary schools. *Health Education* 115(3/4): 322–338.

Glasgow R, Lichtenstein E and Marcus AC (2003) Why don't we see more translation of health promotion research into practice? Rethinking the efficacy-to effectiveness transition. *American Journal of Public Health* 93: 1261–1267.

Griffiths F, Wild A, Harvey J, et al. (2000) The productivity of primary care research networks. *British Journal of General Practice* 50: 913–915.

Hargreaves J, Hatcher A, Strange V, et al. (2010) Group-microfinance and health promotion among the poor: Six-year process evaluation of the Intervention with Microfinance for AIDS and Gender Equity (IMAGE) in rural South Africa. *Health Education Research* 25(1): 27–40.

Hawe P (2015a) Lessons from complex interventions to improve health. *Annual Review of Public Health* 36: 307–323.

Hawe P (2015b) Minimal, negligible and negligent interventions. *Social Science & Medicine* 138: 265–268.

Hawe P, Shiell A and Riley T (2004) Complex interventions: How "out of control" can a randomised controlled trial be? *BMJ* 328: 1561–1563.

Hawe P, Shiell A and Riley T (2009) Theorising interventions as events in systems. *American Journal of Community Psychology* 43(3–4): 267–276.

Hawkins A (2014) The case for experimental design in realist evaluation. *Learning Communities: International Journal of Learning in Social Contexts* 14: 46–59.

Higgins JPT, Altman DG, Gøtzsche PC, et al. (2011) The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* 343: d5928.

Humphreys DK and Eisner MP (2014) Do flexible alcohol trading hours reduce violence? A theory-based natural experiment in alcohol policy. *Social Science & Medicine* 102: 1–9.

Jamal F, Fletcher A, Shakleton N, et al. (2015) The three stages of building and testing mid-level theories in a Realist RCT: A theoretical and methodological case-example. *Trials* 16: 466.

Kirby D (2004) *BDI Logic Models: A Useful Tool for Designing, Strengthening and Evaluating Programs to Reduce Adolescent Sexual Risk-Taking, Pregnancy, HIV and Other STDs*. Scotts Valley, CA: ETR Associates.

Lancaster GA (2015) Pilot and feasibility studies come of age! *Pilot and Feasibility Studies* 1: 1.

Lipsey MW (2009) The primary factors that characterize effective interventions with juvenile offenders: A meta-analytic overview. *Victims and Offenders* 4: 124–147.

Littlecott HJ, Moore G, Moore L, et al. (2014) Psychosocial mediators of change in physical activity in the Welsh national exercise referral scheme: Secondary analysis of a randomised controlled trial. *International Journal of Behavioral Nutrition and Physical Activity* 11: 109.

Macintyre S and Petticrew M (2000) Good intentions and received wisdom are not enough. *Journal of Epidemiology & Community Health* 54: 802–803.

Marchal B, Westhorp G, Wong G, et al. (2013) Realist RCTs of complex interventions - an oxymoron. *Social Science & Medicine* 94: 124–128.

Mayo-Wilson E, Grant S, Hopewell S, et al. (2013) Developing a reporting guideline for social and psychological intervention trials. *Trials* 14: 242.

Merton RK (1968) *Social Theory and Social Structure*. New York: Free Press.
Michie S, Van Stralen MS and West R (2011) The behaviour change wheel: A new method for characterising and designing behaviour change interventions. *Implementation Science* 6: 42.
Moore G, Audrey S, Barker M, et al. (2014) *Process Evaluation of Complex Interventions: Medical Research Council guidance*. MRC Population Health Science Research Network, London.
Moore G, Audrey S, Barker M, et al. (2015) Process evaluation of complex interventions: Medical Research Council guidance. *BMJ* 350: h1258.
Moore G, Moore L, Russell A, et al. (2012) Integration of motivational interviewing into practice in the National Exercise Referral Scheme in Wales: A mixed methods study. *Behavioural and Cognitive Psychotherapy* 40(3): 313–330.
Moore G, Raisanen L, Moore L, et al. (2013) Mixed-method process evaluation of the Welsh National Exercise Referral Scheme. *Health Education* 113(6): 476–501.
Murphy S, Tudor Edwards R, Williams N, et al. (2012) An evaluation of the effectiveness and cost effectiveness of the National Exercise Referral Scheme in Wales, UK: A randomised controlled trial of a public health policy initiative. *Journal of Epidemiology and Community Health* 66: 1082.
Newbury-Birch D, Scott S, O'Donnell A, et al. (2014) A pilot feasibility cluster randomised controlled trial of screening and brief alcohol intervention to prevent hazardous drinking in young people aged 14 15 years in a high school setting (SIPS JR-HIGH). *Public Health Research* 2(6).
Olds D (2016) Building evidence to improve maternal and child health. *The Lancet* 387 (10014): 105–107.
Pawson R (2013) *The Science of Evaluation: A Realist Manifesto*. London: SAGE Publications.
Pawson R and Tilley N (1997) *Realistic Evaluation*. London: SAGE Publications.
Pawson R, Greenhalgh T, Harvey G, et al. (2005) Realist review – a new method of systematic review designed for complex policy interventions. *Journal of Health Services Research and Policy* 10(suppl 1): 21–34.
Pearson M, Brand SL, Quinn C, et al. (2015a) Using realist review to inform intervention development: Methodological illustration and conceptual platform for collaborative care in offender health. *Implementation Science* 10: 134.
Pearson M, Chilton R, Wyatt K, et al. (2015b) Implementing health promotion programmes in schools: A realist systematic review of research and experience in the United Kingdom. *Implementation Science* 10: 149.
Petticrew M (2015) Time to rethink the systematic review catechism? Moving from 'what works' to 'what happens'. *Systematic Reviews* 4(1): 36.
Petticrew M, Rehfuess E, Noyes J, et al. (2013) Synthesizing evidence on complex interventions: How meta-analytical, qualitative, and mixed-method approaches can contribute. *Journal of Clinical Epidemiology* 66: 1230–1243.
Petticrew M, Tugwell P, Kristjannsson E, et al. (2012) Damned if you do, damned if you don't: Subgroup analysis and equity. *Journal of Epidemiology and Community Health* 66: 95–98.
Pronyk P, Hargreaves J, Kim JC, et al. (2006) Effect of a structural intervention for the prevention of intimate-partner violence and HIV in rural South Africa: A cluster randomised trial. *The Lancet* 368 (9551): 1973–1983.
Ragin CC, Drass KA and Davey S (2006) *Fuzzy-Set/Qualitative Comparative Analysis 2.0*. Tucson AZ: Department of Sociology, University of Arizona.
Robling M, Bekkers M-J, Bell K, et al. (2016) Effectiveness of a nurse-led intensive home-visitation programme for first-time teenage mothers (Building Blocks): A pragmatic randomised controlled trial. *The Lancet* 387(10014): 146–155.
Shiell A, Hawe P and Gold L (2008) Complex interventions or complex systems? Implications for health economic evaluation. *BMJ* 336: 1281–1283.
Smith KE, Collin J, Hawkins B, et al. (2016) The pursuit of ignorance. *BMJ* 352: i1446.
Stokols D (2006) Toward a science of transdisciplinary action research. *American Journal of Community Psychology* 38(1): 63–77.

Thomas J, O'Mara-Eves A and Brunton G (2014) Using qualitative comparative analysis (QCA) in systematic reviews of complex interventions: A worked example. *Systematic Reviews* 3: 67.

Thomas P and White A (2001) Increasing research capacity and changing the culture of primary care towards reflective inquiring practice: The experience of the West London Research Network (WeLReN). *Journal of Interprofessional Care* 15(2): 133–139.

Weiss CH (1995). Nothing as practical as good theory: Exploring theory-based evaluation for comprehensive community initiatives for children and families. In: Connell JP, Kubisch AC, Schorr LB and Weiss CH (eds) *New Approaches to Evaluating Community Initiatives: Concepts, Methods, and Contexts*. Washington DC: Aspen Institute, pp. 65–92.

Westhorp G (2012) Using complexity-consistent theory for evaluating complex systems. *Evaluation* 18(4): 405–420.

Westhorp G (2013) Developing complexity consistent theory in a realist investigation. *Evaluation* 19(4): 364–382.

White J, Moore L, Campbell R, et al. (2014) ASSIST+Frank protocol. Available at: http://www.nets.nihr.ac.uk/__data/assets/pdf_file/0017/116027/PRO-12–3060–03.pdf (accessed 24 May 2016).

Whitehead M (2007) A typology of actions to tackle social inequalities in health. *Journal of Epidemiology and Community Health* 61(6): 473–478.

Wong G, Greenhalgh T, Westhorp G, et al. (2013) RAMESES publication standards: Realist syntheses. *BMC Medicine* 11: 21.

Adam Fletcher is a Reader in Social Science and Health at Cardiff University. He has a long standing interest in evaluation methodology, including realist methods.

Farah Jamal was a Research Officer at the UCL Institute of Education. She tragically died in 2016 aged 30 but had already made significant contributions in the field of evaluation.

Graham Moore is a Senior Lecturer at Cardiff University. He led the development and authorship of MRC guidance for process evaluation of complex interventions.

Rhiannon E. Evans is a Research Fellow at Cardiff University. Her work on evaluation methods includes the development of pragmatic formative process evaluation.

Simon Murphy is Professor of Social Interventions and Health at Cardiff University and Director of DECIPHer, a UKCRC centre of excellence for public health research.

Chris Bonell is Professor of Public Health Sociology at the London School of Hygiene and Tropical Medicine. His research focusses on public health and evaluation methods.

Dalkin et al. Implementation Science (2015) 10:49
DOI 10.1186/s13012-015-0237-x

IMPLEMENTATION SCIENCE

DEBATE Open Access

What's in a mechanism? Development of a key concept in realist evaluation

Sonia Michelle Dalkin[1*], Joanne Greenhalgh[1], Diana Jones[2], Bill Cunningham[3] and Monique Lhussier[2]

Abstract

Background: The idea that underlying, generative mechanisms give rise to causal regularities has become a guiding principle across many social and natural science disciplines. A specific form of this enquiry, realist evaluation is gaining momentum in the evaluation of complex social interventions. It focuses on 'what works, how, in which conditions and for whom' using context, mechanism and outcome configurations as opposed to asking whether an intervention 'works'. Realist evaluation can be difficult to codify and requires considerable researcher reflection and creativity. As such there is often confusion when operationalising the method in practice. This article aims to clarify and further develop the concept of mechanism in realist evaluation and in doing so aid the learning of those operationalising the methodology.

Discussion: Using a social science illustration, we argue that disaggregating the concept of mechanism into its constituent parts helps to understand the difference between the resources offered by the intervention and the ways in which this changes the reasoning of participants. This in turn helps to distinguish between a context and mechanism. The notion of mechanisms 'firing' in social science research is explored, with discussions surrounding how this may stifle researchers' realist thinking. We underline the importance of conceptualising mechanisms as operating on a continuum, rather than as an 'on/off' switch.

Summary: The discussions in this article will hopefully progress and operationalise realist methods. This development is likely to occur due to the infancy of the methodology and its recent increased profile and use in social science research. The arguments we present have been tested and are explained throughout the article using a social science illustration, evidencing their usability and value.

Keywords: Realist, Methodology, Palliative care, Realist evaluation, Realist synthesis

Background

The idea that enquiry works by uncovering the underlying, generative mechanisms that give rise to causal regularities has become a guiding principle across many social and natural science disciplines. This article aims to provide a brief description of social mechanisms, mechanisms within evaluation and then specifically mechanisms in realist evaluation. The principles of Pawson and Tilley's [1] conceptualisation of mechanism will then be discussed and operationalised through a reconceptualisation of the Context-Mechanism-Outcome configuration (CMOc) and an understanding of mechanisms on a continuum of activation.

* Correspondence: s.m.dalkin@leeds.ac.uk
[1]University of Leeds, Leeds, UK
Full list of author information is available at the end of the article

Much ado about mechanisms

Social mechanisms

One of the key tenets of realism is the very basic idea that observational evidence alone cannot establish causal uniformities between variables. Rather, it is necessary to explain why the relationships come about; it is necessary to establish what goes on in the system that connects its various inputs and outputs. In this manner, physicists are able fully to understand the relationship between the properties of a gas (as measured by the variables—pressure, temperature and volume) using knowledge about the kinetic action of the constituent molecules. In pharmacology, the term 'mechanism of action' refers to the specific biochemical interaction through which a drug substance acts on the body to generate its curative effect. Programme evaluators do not suppose that CCTV (the intervention) causes a fall in crime rates (the

outcome). It does so, when it does so, by persuading potential perpetrators of increased risks of detection (the mechanism). In all cases, science delves into the 'black box'. In all cases, the mechanism is what generates the observed relationship.

Whilst it is possible to recognise the affinities in explanatory structure across these examples, they also demonstrate that the action of the generative mechanisms is quite different, to such an extent indeed that that they defy a simple, unitary definition of their nature and content. Pawson expands on the applications of generative vs successive conceptualisations of causation elsewhere [2].

Readers of this journal will need no reminding that these paradigms have been debated for many years. Realists see physical and social reality as stratified and emergent. Things that cannot be cast as variables yet are vital to explanation (like kinetic forces, cultural norms and human interpretation or agency) are missing from correlational methods. Causal associations themselves are rarely universal; they are adaptive 'demi-regularities', which are always strongly influenced by setting and context. The original sources for these arguments may be found in Hesse [3], Harré [4], Pawson [2,5], Sayer [6,7], Bhaskar [8], Boudon [9] and Stinchcombe [10].

We acknowledge the further cleft between 'critical realism' and 'scientific realism'. The writings of Bhaskar [8,11] and Pawson [2] serve as a reasonable proxy for these two schools. They differ on the matter of whether social science can create 'closed system' investigations. For Bhaskar, the closed system, experimental control available to the natural scientist is not achievable in social research because of ever-present emergence, that is to say the unique and unceasing human capacity to change the circumstances in which they live. As a 'substitute' for closed system empirical enquiry, he thus proposes the usage of abstract, *a priori* reasoning and the admission of a moral lens through which to critically evaluate human actions ([11], p. 64). Pawson, by contrast, argues much more pragmatically that neither physical science nor social science investigation depends on the achievement of closed systems ([5], p. 67). There are no crucial experiments (most especially Randomised Controlled Trials) which alone furnish us with social laws. But equally, natural science only ever makes slow and imperfect progress in gathering knowledge of the potentially infinite number of contingencies that can shape a physical system. Investigatory closure is always partial. Again, we are presented with rather different visions, the only contradiction occurring when an investigation claims to be *both* normative and scientific.

For Archer [12], collective, constrained decision making is the underlying mechanism that creates all social outcomes. Society is made by but never under the control of human intentions. At any given time, peoples' choices are conditioned by pre-existing social structures and organisations. We are thus externally constrained in our actions but always part of human agency is the choice to attempt to change the initial conditions that bear down on us. These adaptive choices, over time, go on to mould novel structures and changed institutions. Collectively, our present decisions congregate to form new systems, which in their own turn, constrain and enable the choices of the next generation. Society is thus patterned and re-patterned by wilful action, but as Archer reminds us, the causal outcomes never conform to anyone's wishes—even the most powerful.

Most realists would affirm this broad account of the mechanisms of social change, where structures shape actions, which shape structure, which shape actions, and so on. There are, however, some significant differences in where they locate the precise locus of that change. For Bhaskar [8], causal mechanisms sit primarily within the structural component of the social world. They reside in the power and resources that lie with the great institutional forms of society. For other realists, such as Pawson and Tilley [1], mechanisms are identified at the level of human reasoning. Thus, mechanisms can have different meanings depending on the scope of the intended explanation. Structural mechanisms come to the fore if the social scientist is attempting to explain large-scale social transformations. If, however, the researcher is attempting to discover whether a particular fitness programme creates healthier participants, it can be assumed that key outcomes will result from the reasoning and responses of the participants.

Mechanisms in evaluation

This brings us to a consideration of mechanisms in evaluation research; here the focus is on developing an explanation of how a particular programme works through changing the reasoning and responses of participants to bring about a set of intended outcomes. There have been a number of different conceptualisations of mechanism within evaluation. Chen and Rossi [13] were among the first researchers to use the term 'mechanism' and highlight its significance in theory-driven evaluation [14]. In 2005, Chen [15] broadened our understanding of causal mechanisms by identifying two types: mediating and moderating. He defines these as follows:

> *"A mediating causal mechanism is a component of a program that intervenes in the relationship between two other components . . . [while] the second type of causal mechanism—moderating—represents a relationship between program components that is enabled, or conditioned, by a third factor."* (pp. 240–241)

Weiss [16] also reflects on mechanisms, in terms of programme theory. She states that it is important to understand the difference between implementation

theory and programme theory. The earlier can be conceptualised as a logic model, whereas the latter:

> *". . . deals with the mechanisms that intervene between the delivery of program service and the occurrence of outcomes of interest. It focuses on participants' responses to program service. The mechanism of change is not the program service per se but the response that the activities generate."* (p. 46)

As Weiss [16] states, mechanisms are not the programme service but the response it triggers from stakeholders and resulting outcome. For example, Vassilev et al.'s [17] metasynthesis investigated how social networks can make a considerable contribution to improving health outcomes for people with long-term conditions (specifically, type 2 diabetes). They identified three themes which translated into three 'network mechanisms': *network navigation* (identifying and connecting with relevant existing resources in a network), *negotiation within networks* (re-shaping relationships, roles, expectations, means of engagement and communication between network members) and *collective efficacy* (developing a shared perception and capacity to successfully perform behaviour through shared effort, beliefs, influence, perseverance, and objectives). The authors highlight not only resources in these mechanisms but also reasoning; these mechanisms convey the close interdependence between social and psychological processes in long-term conditions management. Furthermore, these network mechanisms are subject to context, as the authors state:

> *"they are shaped by the environments in which they take place which can be enabling or disabling depending on the capacities they offer for carrying out illness management work and supporting behaviours beneficial for people's health."* (p. 10)

Despite the many different conceptualisations, e.g. [9,13-16,18], and applications of mechanisms, e.g. [17,19,20], most in some way have been influenced by the critical realism and scientific realism accounts of causation, e.g. [1,21,22], discussed above. In these schools of thought, mechanisms are usually hidden, sensitive to variations in context and generate outcomes. As Astbury and Leeuw [14] state, mechanisms in realism are:

> *"underlying entities, processes, or structures which operate in particular contexts to generate outcomes of interest."* (p. 368)

We survey this broader terrain as a prelude to focussing on the more specific version of mechanism thinking referred to by Pawson and Tilley that has come to play a key role in the evaluation of social interventions, namely realist evaluation [1], which is the main focus of this article.

Mechanisms in realist evaluation

Within the scientific realism approach, Pawson and Tilley [1] have provided their own conceptualisation of mechanisms; mechanisms are a combination of resources offered by the social programme under study and stakeholders' reasoning in response [1]. They state that mechanisms will only activate in the right conditions, providing a context + mechanism = outcome formula as a guiding principle to realist enquiry [1]. This article sits within the empirical application of realism in the form of realist evaluation and the usage of mechanisms therein. In particular, we make a case for the explicit disaggregation of resources and reasoning in implementation endeavours, to which task we now turn.

The units of analysis within realist evaluation are programme theories—the ideas and assumptions underlying how, why and in what circumstances complex social interventions work. Many readers will by now be very familiar with programme theories expressed as CMOc and with the fact that data collection and analysis in realist evaluation centres on the process of developing, testing and refining CMOc. In the next section of the paper, we propose a development of this formula, which aims to facilitate the study of implementation processes and interventions.

A social science illustrative case study

In order to illustrate our argument in this article and maximise explanatory reach, we draw on empirical data from our realist evaluation of a palliative care Integrated Care Pathway (ICP). The ICP aimed to improve the coordination of care for people in the final year of life by identifying individuals approaching end-of-life, assessing and agreeing how needs and preferences of patients could be met, providing support for families and carers and using Advance Care Planning (ACP) to manage the patients' final illness in order to achieve a 'good' (preference based) death. The ICP comprised a variety of interventions including palliative care registrations, ACP and multidisciplinary team meetings in order to anticipate and plan care for patients with palliative care needs. We evaluated the implementation of the ICP across 14 GP practices in one UK locality using realist evaluation. Five initial programme theories, generated from immersion in the field and literature on ICPs, were tested: (1) the embeddedness of the ICP into GP practices, (2) the registration of palliative care patients, (3) preference discussions and ACP, (4) facilitating difficult conversations and (5) facilitating home deaths. The five refined programme theories were combined to create one overall programme theory of the

whole ICP. This encapsulated the ICP as a translational tool of national policy drivers (such as shared decision making, patient-centred care and proactive care) into local practice.

Using realist evaluation to shed light on how such a complex intervention could work in practice made intuitive sense but proved not to be without operational challenges. These have been echoed by other realist researchers [23-25] and have prompted the writing of this paper.

This paper has two main aims:

- To make a case for the explicit disaggregation of resources and reasoning within mechanisms;
- To reiterate the need for nuance in considering whether mechanisms fire in a dual on/off mode.

Discussion

Disaggregating mechanisms into resources and reasoning

1 The concern

Realists posit that exposing not only the mechanisms of change in an intervention but more importantly their relationship to the context of their implementation is key to the evaluation of complex programmes [20,26]. However, deciding whether aspects within an intervention implementation process in a realist project contribute contextually or mechanistically to the overall explanatory endeavour has become the realist researcher's quandary [14,23,27]. Like these authors, we encountered challenges in distinguishing between context and mechanism in our evaluation of the ICP and were cognisant of the need not to conflate programme strategy (the intervention) with mechanism. We concur with Jagosh et al. [23], who note how it is not always as straightforward as might be assumed to map the complexities of the transformation process and the multiple systems within which it operates onto the C + M = O formula. Arguably, outcomes can be identified with most ease; they are observed or measured or at least aimed at with a degree of clarity. Although the distinction between resources and reasoning is used in Pawson and Tilley's seminal work [1], their relative importance in understanding mechanisms is often understated. Consequently, researchers often emphasise one at the expense of the other, under the banner of mechanism [25]. To address this, we offer the solution below.

2 Our way forward

Building on the original work of Pawson and Tilley [1], we would like to propose an alternative operationalisation of the CMOc formula:

Intervention resources are introduced in a context, in a way that enhances a change in reasoning. This alters the behaviour of participants, which leads to outcomes.

The revised formula therefore reads:

$$\mathbf{M(Resources) + C \rightarrow M(Reasoning) = O}$$

Resources and reasoning are mutually constitutive of a mechanism, but explicitly disaggregating them can help operationalise the difference between a mechanism and a context. Although resource and reasoning are made explicit in the seminal work of Pawson and Tilley [1], they have often not been referred to explicitly in subsequent research. In our own study, through using this formula, it became clearer whether data contributed contextually or mechanistically, as we could identify mechanism components (resource and reasoning) which are different to contexts. Figure 1 illustrates how we have presented the new formula diagrammatically in the ICP study. Through trial and error, it became clear that the original formula could be built upon, hence the new formula which disaggregates resource and reasoning, placing 'context' in between. However, this is not to be confused with just using resources without reasoning—they must always come as a pair. It is important to note here that this new formula is only an extension of the original heuristic developed by Pawson and Tilley [1]. This new formula does not aim to re-draw the full sequence of causation but to modify the basic heuristic to aid operationalisation of realist approaches.

Differentiating between resource (the component introduced in a context) and reasoning therefore helps distinguish between relevant context and mechanism. Identifying the resource is contingent on the purpose of the study, and identifying the reasoning avoids the issue of conflating programme strategy (resource) with mechanism.

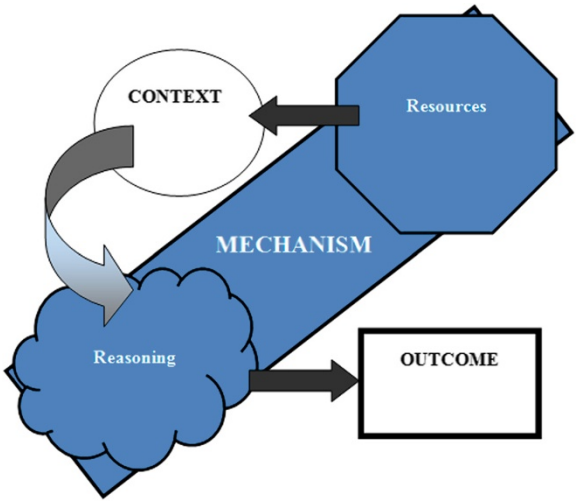


Figure 1 A CMOc framework.

3 The social science illustration

In the palliative care ICP study, an outcome pattern was observed that practices identified and placed fewer palliative patients with non-cancer illnesses on their palliative care registers, in comparison to those with cancer illnesses. This was common across all 14 practices studied and was particularly noticeable for patients residing in care homes, where many older adults have non-cancer illnesses. Patients with non-cancer illnesses have unpredictable illness trajectories, meaning that registering this patient group is challenging for health care professionals, as a period of significant decline can be followed by substantial improvement, despite a downward trend in wellness [28,29]. Comparatively, this is not the case with cancer diagnoses as often there is a specific diagnosis and steady illness trajectory. We aimed to generate a CMOc to explain why there were less palliative care registrations of patients with non-cancer illnesses than cancer patients (outcome). In attempting to formulate the configuration, we were uncertain whether the context was the unpredictable illness trajectories of older adults without a cancer diagnosis, or care homes in general or the palliative care register being difficult to use with non-cancer patients. Breaking down the C + M = O formula to include resource and reasoning using the new formula, M (resource) + C → M (reasoning) = O, helped in deciphering the context from the mechanism. The use of the new formula diagram (Figure 2) also helped in configuring the whole CMOc. Figure 2 displays the novel way in which the new formula should be represented diagrammatically. Through using the new formula and associated diagram, it became clear that the resource was the palliative care register which, when used with older adults who had unpredictable illness trajectories (context), resulted in anxiety in registering these patients (reasoning), which meant that less older patients in care homes were registered (outcome) (Figure 2). Through understanding that resources were introduced into pre-existing contexts in a way that altered the participants' reasoning, it becomes easier to explain the differential registration numbers (outcome).

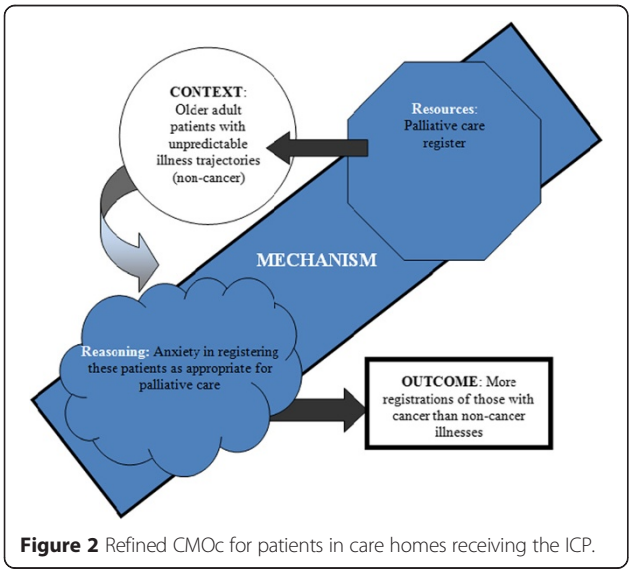


Figure 2 Refined CMOc for patients in care homes receiving the ICP.

Disaggregating resources and reasoning encourages researchers to consider both concepts, rather than privileging one at the expense of the other and will contribute significantly to the explanatory endeavour of the realist researcher. It is important to understand the new formula (M (Resource) + C → M (Reasoning) = O) highlights that resources must be introduced into a pre-existing context, which in collaboration induces an individual's reasoning, leading to an outcome. Distinguishing the resources that are introduced into contexts from the reasoning this generates can provide both an operational and a conceptual clarification of mechanism. It can enable researchers to clearly understand the role of context in triggering mechanisms, thus developing their explanation of how interventions work. We now turn to interrogate the notion of mechanisms being 'triggered' in the next section of the paper.

A case for continuums of activation in reasoning

1 The concern

A separate but related difficulty encountered when using mechanisms in social science research is the notion that mechanisms are often said to 'fire', 'trigger' or 'modify' in context to create an outcome [1,30-32]. Pawson and Tilley [1] use the much referenced gun powder analogy to explain this. When a spark is introduced to gun powder, the chemical composition of gun powder (mechanism) results in an explosion (outcome). However, there are no explosions if the context is not right—damp conditions, insufficient powder, not adequately compact, no oxygen present, duration of heat applied is too short (context). Thus it purports that causal outcomes follow from mechanisms acting in contexts; this is the base from which all realist explanation builds. Most complex social interventions involve stakeholders' volition (reasoning). As Pawson [33] states, "much more than in any other type of social programme, interpersonal relationships between stakeholders embody the intervention" [33]. We found it difficult to apply the firing analogy to interventions where human volition is entwined in the intervention. Reasoning in these cases is rarely activated via an on/off switch, triggered in favourable contextual conditions. Instead, activation operates along a continuum similar to the light created by a 'dimmer switch', where intensity varies in line with an ever evolving context. Our experience suggests that researchers are often enabled to develop their realist thinking further when this myth of on/off reasoning is dispelled. The metaphor of the dimmer switch accommodates the activation of new volition as well as the idea of continuums of activation.

2 Our way forward

Conceptualising volition as happening in a binary 'firing'/'not firing' fashion masks a continuum of activation which can have more explanatory value in understanding how interventions work. There are varying degrees to which an individual can feel confident, angry or mistrustful, leading in turn to a gradation of outcomes.

3 The social science illustration

In our evaluation of the ICP, we observed that the volition of health care professionals was always on a continuum. Health care professionals felt anxious when registering older adults with an illness other than cancer, as the trajectory of such illnesses is so unpredictable (Figure 2). Health care professionals could not predict patients' decline, did not wish to over populate their palliative care registers and were worried about registering patients who seemed relatively well but could decline quickly. Furthermore, once a decline in health begins in older adults with non-cancer illnesses, it can be very rapid and thus end-of-life care is implemented quickly and is often unplanned, which can result in a death that does not adhere to patient preferences. The anxiety of health care professionals working with palliative non-cancer patients was evident, yet this anxiety did not switch on and off, it developed over time, as patients' illnesses progressed. It also differed between health care professionals; those with more experience of working with patients with non-cancer disease had less anxiety about registering them. Thus the reasoning of having anxiety was on a continuum for health care professionals using the palliative care register. There is a variation in the amount of anxiety a health care professional will feel when registering a patient with a non-cancer illness, it is not dichotomised; the degree to which this is felt is combined with a facilitative context and appropriate resource. This should lead to a more appropriate use of the palliative care register.

Summary

This paper aimed to help the operationalisation of the C + M = O formula, through (1) a disaggregation of the mechanism resource and mechanism reasoning and (2) a conceptualisation of activation continuums, rather than a binary trigger. The solutions proposed in this article will enable a clearer application of realist evaluation to understanding how complex interventions are implemented. We have already found some evidence to support this argument by applying it in our own teaching and workshops. For example, the 'workability' of this framework has been tested with researchers at the beginning of their realist journey in a realist summer school at the Centre for Advancement in Realist Evaluation and Synthesis (CARES), University of Liverpool. Course participants found it useful to guide their realist learning, understand the method further and clarify the differences between mechanism and context, and resources and reasoning.

We hope that this article furthers the discussions on the operationalisation of realist theory development in a way that, in particular, helps novice realist researchers to embrace and in turn develop the methodology. The authors would welcome testing of the methodological refinements discussed throughout this article by other researchers across a wide range of fields, with such testing aiding further developments.

Abbreviation

CMOc: Context Mechanism Outcome configuration.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SMD conceived and drafted the article. ML, JG, DJ and BC commented on and revised the article. All authors read and approved the final manuscript.

Authors' information

SMD conceived this article whilst writing her PhD using realist evaluation. She is now a Research Fellow using realist methods at the University of Leeds, working with JG. ML, DJ and BC supervised SMD's PhD.

Acknowledgements

Acknowledgements to Northumbria University and NHS North of Tyne for joint funding the PhD this paper is based on. We would also like to acknowledge Professor Ray Pawson for his methodological guidance and expertise throughout the refinement of this article. Useful comments were also received from three referees.

Author details

[1]University of Leeds, Leeds, UK. [2]Northumbria University, Newcastle Upon Tyne, UK. [3]Hadrian Primary Care Alliance, Newcastle Upon Tyne, UK.

Received: 17 November 2014 Accepted: 24 March 2015
Published online: 16 April 2015

References

1. Pawson R, Tilley N. Realistic evaluation. London: SAGE; 1997.
2. Pawson R. A measure for measures: a manifesto for empirical sociology. London: Routledge; 1989.
3. Hesse M. The structure of scientific inference. Oakland, CA: University of California Press; 1974.
4. Harré R. The philosophy of science: an introductory survey. London: Oxford University Press; 1972.
5. Pawson R. The science of evaluation: a realist manifesto. London: SAGE; 2013.
6. Sayer A. Realism and social science. London: Sage; 2000.
7. Sayer A. Method in social science: a realist approach. London: Hutchinson; 1984.
8. Bhaskar R. A realist theory of science. 2nd ed. Brighton: Harvester Press; 1978.
9. Boudon R. Social mechanisms without black boxes. In: Hedström P, Swedberg R, editors. Social mechanisms: an analytical approach to social theory. UK: Cambridge University Press Cambridge; 1998.
10. Stinchcombe A. The conditions of fruitfulness of theorizing about mechanisms in social science. Philos Soc Sci. 1991;21(3):367–88.
11. Bhaskar R. The possibility of naturalism: a philosophical critique of the contemporary human sciences. Brighton: Harvester Press; 1979.
12. Archer M. Realist social theory: the morphogenic approach. Cambridge: Cambridge University Press; 1995.
13. Chen H, Rossi P. The theory-driven approach to validity. Eval Program Plann. 1987;10:95–103.
14. Astbury B, Leeuw F. Unpacking black boxes: mechanisms and theory building in evaluation. Am J Eval. 2010;31(3):363–81.
15. Chen H. Practical program evaluation. Thousand Oaks, CA: SAGE; 2005.

16. Weiss C. Theory-based evaluation: past, present, and future. New directions for evaluation. San Francisco, CA: Jossey-Bass; 1997.
17. Vassilev I, Rogers A, Kennedy A, Koetsenruijter J. The influence of social networks on self-management support: a metasynthesis. BMC Public Health. 2014;14(719):1–12.
18. Hedstom P, Swedberg R. Social mechanisms: an analytical approach to social theory. Cambridge: Cambridge University Press; 1998.
19. Thoits P. Mechanisms linking social ties and support to physical and mental health. J Health Soc Behav. 2011;52(2):145–61.
20. Greenhalgh T, Humphrey C, Hughes J, Macfarlane F, Butler C, Pawson R. How do you modernize a health service? A realist evaluation of whole-scale transformation in London. Milbank Q. 2009;87(2):391–416.
21. George A, Bennett A. Case studies and theory development in the social sciences. Cambridge, MA: MIT Press; 2004.
22. Henry G, Julnes G, Mark M. Realist evaluation: an emerging theory in support of practice. New directions for program evaluation. San Francisco, CA: Jossey-Bass; 1998.
23. Jagosh J, Pluye P, Wong G, Cargo M, Salsberg J, Bush PL, et al. Critical reflections on realist review: insights from customizing the methodology to the needs of participatory research assessment. Res Synth Methods. 2013;5(2):131–41.
24. Salter K, Kothari A. Using realist evaluation to open the black box of knowledge translation: a state-of-the-art review. Implement Sci. 2014;9(115):1–14.
25. Pawson R, Manzano-Santaella A. A realist diagnostic workshop. Evaluation. 2012;18:176–91.
26. Berwick D. The science of improvement. JAMA. 2008;299(10):1182–4.
27. Marchal B, van Belle S, van Olmen J, Hoerée T, Kegels G. Is realist evaluation keeping its promise? A literature review of methodological practice in health systems research. Evaluation. 2012;18(192):192–212.
28. Murtagh F, Preston M, Higginson I. Patterns of dying: palliative care for non-malignant disease. Clin Med. 2004;4:39–44.
29. Murray S, Kendall M, Boyd K, Sheikh A. Illness trajectories and palliative care. Br Med J. 2005;330:1007–11.
30. Jagosh J, Macaulay A, Pluye P, Salsburg J, Bush PL, Henderson J, et al. Uncovering the benefits of participatory research: implications of a realist review for health research and practice. Milbank Q. 2012;90(2):311–46.
31. Wilson V, McCormack B. Critical realism as emancipatory action: the case for realistic evaluation in practice development. Nurs Philos. 2006;7(1):45–57.
32. Wong G, Greenhalgh T, Westhorp G, Buckingham J, Pawson R. RAMESES publication standards: realist syntheses. BMC Med. 2013;11(21):1–14.
33. Pawson R. Digging for nuggets: how 'bad' research can yield 'good' evidence. Int J Soc Res Methodol. 2006;9(2):127–42.

What Is Population Health Intervention Research?

Penelope Hawe, PhD,[1] Louise Potvin, PhD[2]

ABSTRACT

Population-level health interventions are policies or programs that shift the distribution of health risk by addressing the underlying social, economic and environmental conditions. These interventions might be programs or policies designed and developed in the health sector, but they are more likely to be in sectors elsewhere, such as education, housing or employment. Population health intervention research attempts to capture the value and differential effect of these interventions, the processes by which they bring about change and the contexts within which they work best. In health research, unhelpful distinctions maintained in the past between research and evaluation have retarded the development of knowledge and led to patchy evidence about policies and programs. Myths about what can and cannot be achieved within community-level intervention research have similarly held the field back. The pathway forward integrates systematic inquiry approaches from a variety of disciplines.

Key words: Evaluation; population health intervention research; evidence-based practice; intervention research; population health

La traduction du résumé se trouve à la fin de l'article.

Can J Public Health 2009;100(1):I8-I14.

There is an increasing move worldwide to shift the emphasis of population health research away from purely descriptive and analytic studies and towards the study of interventions to reduce health problems and reduce health inequities.[1] This requires an appreciation of the best that we have learned from the diverse settings in which both health and social scientists have been working. As far as possible we call for an integration of that learning to assist in the development of the relatively new overarching field of population health intervention research.

This paper outlines the practice of intervention research in population health. We draw attention to a number of features that mark problematic and unnecessary distinctions between intervention research and evaluation research, arguing that these fields comprise very similar research practice and orientation. We discuss the skill sets involved in this type of research and end by describing and debugging common myths with regard to intervention research.

Defining intervention research

The definition of intervention draws from its Latin roots, *venire*, meaning to come and *inter*, meaning between, drawing attention from the outset that to intervene literally means to come in between, to disturb the "natural" order of things or a foreseeable sequence of events. If we characterize descriptive or analytic research in population health as seeking to understand phenomena, then intervention research is about testing those understandings by the act of intervention in the causal mechanisms under investigation. It is also about learning from the actions implemented to address those phenomena in order to improve our practice. The iconic figure of John Snow removing the handle of the Southwark and Vauxhall Company water pumps that he suspected were responsible for the London cholera outbreak[2] is a dramatic example.

The Population Health Intervention Research Initiative for Canada (PHIRIC) defines population health intervention research thus:

> *Population health intervention research involves the use of scientific methods to produce knowledge about policy and program interventions that operate within or outside of the health sector and have the potential to impact health at the population level.*[3]

We use the term "population health" in the way it is used in Canada to refer to the science underpinning the practice of public health and understandings about health that come only from an appreciation of how health is generated in populations. However, we recognize that in many countries the term "population health" is less used, and hence here "population health research" and "public health research" can be taken to mean the same thing.

The definition refers to the use of scientific methods that have informed the development of many disciplines. In the case of public health, the original critical scientific developments were about social statistics and virology.[4,5] In public health the tradition of intervention research is closely linked to that of experimental medicine, which goes back to the pioneer work of Claude Bernard. The principles of experimental medicine as proposed by Bernard are to systematically examine and, if possible, isolate the physiological consequences of actions undertaken in response to ill health and to try to reproduce those consequences under various conditions. For Bernard, as for most scientists of his time (1870s), causality and scientific laws are only possible through the decomposition of the mechanism, understood as the sequence of events that produces an effect. Altering the outcome of a sequence in a predicted direction constitutes evidence of the truth or validity of the scientific proposition. Today, scientific methods from a variety of disciplines, including the social sciences, are included in the evaluator's toolbox.

The PHIRIC definition also points to interventions both inside and outside the health sector and is neutral on the intentionality

Author Affiliations

1. Population Health Intervention Research Centre, University of Calgary
2. Léa-Roback Research Centre on Social Health Inequality, University of Montreal

Correspondence and reprint requests: P. Hawe, Population Health Intervention Research Centre, University of Calgary, G012, 3330 Hospital Drive NW, Calgary, AB T2N 4N1, Tel: 403-210-9383, Fax: 403-220-7272, E-mail: phawe@ucalgary.ca

Acknowledgements: Many thanks to Adria Rose and Erica Di Ruggiero for helpful comments on an earlier draft. Penelope Hawe is a Health Scientist of the Alberta Heritage Foundation for Medical Research and the Markin Chair in Health and Society. Louise Potvin is the holder of the CHSRF-CIHR Chair on Community Approaches and Health Inequality.

of those events. If we confined ourselves to actions within the health sector intended to improve health we would miss a great deal. The definition reflects an interest in the social determinants of health – economic policy, education policy and environment policy. Actions in sectors outside health, designed for purposes other than health, are often studied by people within health as "natural experiments", e.g., road construction, factory closures, food market openings. Studies of the impact of such events are included in the PHIRIC definition of population health intervention research along with ongoing practices and policies in sectors other than health that might affect population health. Evidence about population health impact has been successful so far in changing practices in the motor vehicle construction industry,[6] the food and beverage industry[7] and the petroleum industry.[8]

The final aspect of the definition is "impact at the population level". The Canadian Institutes of Health Research, Institute of Population and Public Health, reminds us that this does not simply mean improving health or reducing health risks but, rather, involves interventions intended to change the conditions of risk in order to alter the distribution of health risk[9] in keeping with the ideas of Geoffrey Rose.[10] To be truly effective, a population health intervention should be reducing risk exposure in successive cohorts of people within the setting(s) under investigation.

Is intervention research the same as implementation research?

Systematic observation built up around the roll-out of programs and policies as they are implemented in order to appreciate reach, context-level adaptation and effects has a strong tradition in the field of public administration, where it tends to be called "implementation research".[11-13] But what is more associated with the phrase "intervention research" in the health field is the notion that its primary purpose is to test a hypothesis or causal pathway. Hence, attribution of effect to that intervention is a primary driver of the study design in intervention research, as its origin in the 19th century underlines. Note that some health researchers have reserved the term "implementation research" for a phase of work that follows the demonstration of a program's or policy's effects.[14] This implementation research phase is designed to elucidate more understanding about the process of a program that has already shown its effectiveness in a demonstration trial.[14] This idea is often seen in clinical settings.[15] However, others have argued that systematic observation of and improvement in process and implementation should precede the measurement of effects. Indeed to not do so might diminish the chance of a new intervention achieving its effects.[16]

The stepped-wedge cluster-randomized trial design has evolved quite recently for situations in which there is high demand for a policy or program of unknown effectiveness and unlikely harm but insufficient resources for the program or policy to be uniformly provided initially.[17] Policy-makers are often more likely to be persuaded to adopt this staged, randomized roll-out design than a traditional cluster-randomized trial. The stepped-wedge design allows effectiveness to be assessed optimally while local demand is served. From a traditional researcher's perspective the stepped-wedge design remains an effectiveness trial and a classic case of intervention research. For policy-makers it is perhaps seen more as a progressive introduction of a policy or program, with more data-gathering around it than they are perhaps used to. Dual perspectives are important and not incongruent.

We take the view that all systematic inquiry and learning from observing an intervention's process or implementation, impact or outcome is encompassed in the term "intervention research". Terms may always be used differently in different fields, but better research overall will derive from understanding the contributions that have come from different vantage points. In this sense, rather than insisting on any particular language, it is better to pin down the purpose of the research (e.g., testing effectiveness, elucidating the process of action, documenting variations in implementation, tracking reach into populations with highest needs) and to ensure that the methods appropriately match that purpose.

Is intervention research the same as evaluation research?

Evaluation involves making judgements about the worth or value of something.[18] Evaluation research is about the use of scientific methods for that purpose.[19] The focus of enquiry is to interpret an action and make a pronouncement about it, according to values or standards that are pre-set (and usually enshrined in the goals and/or objectives of the action). Evaluation is usually broken down into components named variously but usually encompassing process evaluation (how well an intervention is delivered, whether it reaches the intended target group), impact evaluation (immediate effects) and outcome evaluation (subsequent or longer-term effects). The goals or objectives are enshrined within the sample size calculation that is required for quantitative studies in impact and outcome evaluation. The amount of desired important change is pre-specified.

Evaluation research and population health intervention research encompass many of the same activities and methods. All evaluation research in population health is population health intervention research, but not all population health intervention research is evaluation research. This is because some population health intervention research assesses the health impact of policies and programs in sectors outside of health. Because these policies and programs were not designed with a health outcome or objective in mind they do not conform to the definition of evaluation research in the sense that the criteria for interpreting the health impact are not preset. However, differences between the two fields of work have emerged in practice, some of which are listed in Table 1.

These differences mark a division in the culture of the two areas, in an everyday sense, that has some worrying features. They represent issues we need to address in Canada if we are to gain fully from all the work being undertaken to understand ways to improve health at a population level.

First, there is a tendency to see evaluation as "not research". In many health regions in Canada, evaluation projects are not sent for ethical review and therefore the investigation methods proposed are not scrutinized externally. This may compromise quality.

Second, important questions about intervention effectiveness may be being pursued under the guise of evaluation research and hence commissioned with insufficient resources to pursue answers adequately. This contributes to a poverty of evidence on important issues. For example, the US Task Force on Community Preventive Services has recently lamented that in 50% of the interventions

Table 1. Common Differences That Have Arisen between Intervention Research and Evaluation Research

Intervention Research	Evaluation Research
Intervention is often initiated by the researcher, although it may be designed in collaboration with practitioners.	Intervention under investigation is usually designed by practitioners or agencies.
Funded by a research grant.	Funded by resources within the commissioning agency.
Budgeted according to the information required and the cost to produce that information.	Funded as percentage of the cost of the program, e.g., at an arbitrary level of 10%.
Results are destined for the public domain, e.g., in peer-reviewed journals.	Results may be restricted to an internal report by contract agreements.
Usually focused on assessment of intervention outcomes, assisted by a large budget for primary data collection. May also include assessment of process and mechanism of action.	Smaller budgets frequently limit enquiry to secondary data sources in relation to outcome or restrict the evaluation questions to matters of intervention process, reach or consumer satisfaction.
Requires ethics approval.	Ethics approval not routinely sought.

reviewed there was insufficient research evidence to make any practice recommendations.[20] This is probably not because a wide variety of practices have not been investigated; it is more likely that what has been investigated has not produced evidence that the Task Force considers worthwhile.

A typical scenario comes from the World Health Organization's (WHO) Safe Communities project. There are more than 80 such projects worldwide, which are designed to mobilize and involve communities in reducing injuries. However, only seven of those projects have undertaken controlled evaluations using objective sources of injury data, and only two have been shown to be effective.[21] The dearth of evidence most probably arises because evaluations of Safe Communities projects typically are commissioned by local agencies, with budgets insufficient to employ more than one person full time. They therefore tend to address questions about who is involved in the project, what people think about it, what activities have been conducted and whether inter-sectoral collaboration increased, as described in a case study funded by a state health department in Australia, for example.[22]

Finally, the divide between evaluation and intervention research has meant that a different body of knowledge has evolved to serve each professional field, and important opportunities for cross-development have been lost or delayed. For example, journals such as *Evaluation*, *Evaluation Review*, *New Directions for Evaluation, Evaluation and Program Planning* and *Evaluation and the Health Professions* have published numerous studies about implementation assessment, the importance of context assessment and theories of change processes for a decade or more. Yet it has only been comparatively recently that these notions have been given prominence in the field of evidence generation in public health.[23] One reason for PHIRIC to bring these two fields closer together is to ensure that such misadventure does not persist. Some of the "great failures" in population health intervention research[24] can be attributed to issues that experienced evaluation researchers would have detected earlier. These include issues like inadequate intervention implementation, failure to stage the design of the research to the stage of the intervention's development and/or inadequate program theory. These are domains that health promotion evaluators have long been encouraged to examine systematically at the outset of study design during the process of evaluability assessment.[16] We note, for example, that the failure of the Stanford Heart Disease Prevention Program was predicted at the start by those who argued in journals published at the time that its community-based change logic (theory) was weak.[25] A formal evaluability assessment of the intervention might have held the investigators' decisions around this up to greater scrutiny and debate.

The skill set of an intervention researcher

As the task of intervention research is laid out, it becomes apparent that the skill set to accomplish the task is complex. Technical competence in empirical enquiry is vital and, given the breadth of tasks – e.g., study design, questionnaire design, interview design, data management, data extraction, statistical analysis, qualitative data analysis, economic evaluation and economic modelling – may require the resources of a multidisciplinary team. However, any or all of these skills are part of the regular repertoire of any population health researcher. Because population health interventions are designed to address social conditions that determine risk, a good intervention researcher must have additional skills, including those that allow him or her to play a strategic role in the development and uptake of high-quality interventions (assuming here that the intervention research is real time and not historical, using secondary data sets).

In the first instance, researchers must be able to theorize about change dynamics. Intervention research is about transformation processes. Thus a researcher might need to look for more things, different things or different things in different ways, than if he/she were doing a descriptive or an analytical study. Investigators who have had to deal with the ramifications of interventions (side effects, unintended effects)[26] and the possibility that interventions could cause harm have been led to theorize at multiple levels.[27] McLeroy reminds us that the "theory of the problem" and the "theory of the solution" are not the same.[28] Some of the modest or negative findings in population health intervention research might be attributable to investigators, frustrated with their work in documenting the problems, trying their hand at intervention design and intervention research without a thorough appreciation of the demands of intervention theory and practice. Unfortunately, as a consequence, it may be hard to get policy-makers to reinvest in interventions and intervention research in areas where previous investigation has failed. Put crudely, the baby easily gets thrown out with the bathwater.

Skills in communication, policy and social analysis are vital.[29] Research has to be meaningful and convenient to the people and organizations with whom the researcher is working. Intervention research is about contributing directly to the implementation of actions to improve the population's health. Yet, too often researchers have been accused of designing and testing interventions that no one would be able to implement in real life, ignoring policy-makers' needs.[30] The field of utilization-focused evaluation is helpful here in increasing researchers' sensitivity to stakeholder or end-user needs.[31] Additionally, researchers need to gain the support of practitioners. This can be difficult. Not only does research

Table 2. Examples of the Diversity of Intervention Studies to Address Various Evaluation Questions

Evaluation Question of Interest	Authors	Examples
RELEVANCE: *How relevant is the program to targets of change?*	Bisset et al., 2004[38]	Examines how decisions about the goal and mission of a community diabetes prevention program were informed by prevalence studies conducted in the community.
	Baker et al., 2007[39]	Shows how a community prevention program to reduce childhood obesity was designed on the basis of community asset mapping and led to community engagement in the program.
COHERENCE: *How does the theory of change underlying the program relate to the theory of the problem?*	Hawe & Stickney, 1997[40]	Describes how, despite good will, an intersectoral food policy committee was lacking a mechanism to successfully pursue its goals.
	Levesque et al., 2005[41]	Examines the correspondence between the activities implemented in a community diabetes prevention program and the principles of the socio-ecological approach to health promotion.
RESPONSIVENESS: *How is program implementation responsive to local conditions?*	Ho et al., 2006[42]	Examines how the local conditions prevailing in remote communities were related to changes in the implementation of a First-Nations Diabetes Prevention Program that had been successfully evaluated.
	Corrigan et al., 2006[43]	Describes how the use of qualitative methods helped improve the fit between implementation variations in a randomized trial of secondary prevention and local needs and conditions
ACHIEVEMENTS: *What did program activities and services achieve?*	Wickizer et al., 1998[44]	Identified the critical factors for successful implementation of a community health promotion initiative in 11 communities randomly assigned to receive program grants.
	Cooke et al., 2007[45]	Examines the changes in aggressive and related behaviors as well as in discipline referrals following the successful implementation of a violence prevention program in six schools.
RESULTS/IMPACT: *With which changes in local conditions was the program associated?*	O'Loughlin et al., 1999[46]	Quasi-experimental study showing that although several implementation indicators revealed a high level of program penetration in the community, there was no improvement in health and behavioural indicators.
	Wagenaar et al., 2006[47]	Quasi-experimental study showing positive trends in many indicators in the 10 US States where the Reducing Underage Drinking through Coalition Project funded coalitions designed to change policy and normative environments.

take up practitioners' time but it can also attract resources that would have otherwise been spent on the intervention itself. Careful navigation is required when researcher and practitioner interests do not coincide at the outset.

There may be covert as well as overt reasons for programs and services, and insensitivity on the part of the researcher can fail to recognize this. One example comes from DARE, Drug Abuse Resistance Education in North America. This is a school-based substance abuse prevention program that has been delivered to more than 33 million school children at an annual cost of $0.75 billion. Repeated evaluations have shown that it does not prevent substance abuse.[32] In 2001, a $13.7 million program renovation was undertaken, but there has still been no evidence that the new DARE is effective.[33] However, a recent qualitative analysis has suggested that past researchers have possibly missed the point of the program. Its primary benefit is perceptual – principals and teachers like having police in schools making contact with children and youth.[34] On this basis, schools may wish the program to continue in spite of its failure to prevent drug use. From a population health perspective, there may be cheaper ways of building school collaborations with the police that could allow the bulk of DARE costs to be diverted to more effective programs against drug abuse. The point is that more astuteness on the policy analysis side might have anticipated this finding many years previously.

Common myths in intervention research

In practice, intervention research has often tended to be associated with investigator-driven studies, and evaluation research has been associated with studies commissioned by or conducted with the research users, consumers or decision-makers. This means that a particular profile, or image, of intervention research has arisen that needs to be interrogated.

Myth 1. Intervention Research Is Just About Intervention Effect

We have defined intervention research in a way that emphasizes its role in understanding causal mechanisms, but showing that something is effective is only part of the task. Intervention research is about all parts of the process of designing and testing solutions to problems and about getting solutions into place. It can involve process evaluation of interventions (assessing reach, implementation, satisfaction of participants, quality). It can involve assessment of how interventions adjust to different contexts.[35] It can extend to examinations of how interventions are sustained over time or become embedded in the host institutions.[36] It includes diffusion research or understanding of how interventions are spread to new sites.[37] The WHO Task Force on Health Promotion Evaluation proposed five questions as points of entry for enquiring about an intervention with these multiple aspects in mind.[37] Table 2 presents examples of intervention studies conducted in relation to each of those five questions.

Myth 2. Interventions Designed and Implemented With Communities Should Not Be Called Intervention Research

The field of population and public health is interdisciplinary, eclectic and contested. In community-based intervention research some unhelpful schisms have grown up between studies primarily designed and controlled by researchers and those that are driven by communities. We believe that both types of interventions must be accountable for their logic, values and outcomes.

The whole spectrum of intervention research should be supported, from those interventions driven by hypotheses formulated in academia to those in which interventions are designed and implemented by local actors.[48] However, we stop short of the suggestion that, with respect to communities, the term "intervention" research be dropped in favour of terms like "community development" or "community-based action". The latter terms frame a tra-

dition that is highly respected and characterized by particular ways of working.[49] We are not suggesting that these terms be replaced, but we suggest that it may be appropriate to use the term "intervention research" in conjunction with them when data are being collected by researchers, because the intervention terminology enshrines a dynamic that is important to remember: that of disturbing the regular order of things. A focus on disturbing, interrupting or changing an expected sequence of events draws attention to the ethical issues involved, the relationship between the researcher and the researched and the duty of care enshrined in the relationship.[50] These are neglected issues in population health research, which we believe could become even more neglected if inadvertently hidden by language that disguises the duties and responsibility of the researcher. When researchers become actors in local events, as opposed to being merely observers, many of us find ourselves untrained and unprepared. Preserving a language that alerts us to the special nature of this role is precautionary and vital.

Myth 3. Intervention Research Is Only About Controlled Trials

A lot of intervention research *has* been about controlled trials in schools, worksites and communities, but it does not have to be. Many different types of study design can be used to build acceptable evidence in public health, although some scenarios are more desirable than others when it comes to making causal inferences.[51] Important work is advancing in the use of time series designs to illuminate the impact of policy[52,53] and in the use of observational methods to investigate the relationship between program implementation conditions and impact. There has also been an expansion of community-based participatory research, which has been critical for addressing social determinants of health in communities.[54,55]

Myth 4. Intervention Research Is About Controlled Interventions

There is no reason for all interventions to be as tightly controlled as many investigators have imagined.[56] Indeed it has been observed that the reason so many interventions in schools, worksites and communities have failed may be because investigators have overcontrolled the form of their interventions in the mistaken belief that this is a design requirement of randomized controlled trials.[56]

An alternative way of thinking about standardization has been proposed that can liberate the randomized controlled design and aid its use in more contexts.[56] The key issue is that interventions have to be well theorized and recognizable, so that the evaluation is valid and so that another person could replicate the intervention in another place. The essence of an intervention might be a process or set of functions.[56] This type of intervention follows recognizable principles (a standard function, like organizational development or community development) but necessarily takes a different form from place to place and in that sense owes its effectiveness to how it is tailored to context.[56] Alternatively, the intervention could be fixed or standard in form, like a leaflet based on the health belief model, which draws its benefit from being sufficiently effective overall, even though it is largely immune to local context and not effective in every place.[57] The point is that interventions standardized by form *or* standardized by function can be evaluated meaningfully in randomized controlled trials.[56] Theorizing this at the outset is part of trial design.

The myth that interventions have to be tightly controlled in terms of form unfortunately continues to be promulgated by researchers who use analogies from drug trials to explain the efficacy of community health interventions.[14,58] In these analogies, the most efficacious interventions are framed as those designed by universities or expert authorities of some type. The effectiveness of interventions is then considered to be progressively diluted by the transfer of these technologies from the academy into the hands of local community practitioners.[14,58] The alternative view defines efficacy as starting with interventions designed or shaped by communities and practitioners. These may intersect with universities to the extent that such relationships may be required to strengthen intervention theory and to gather convincing evidence that such interventions work. These alternative views recognize the agency of the practitioners and the capacity of communities to foresee and shape the types of intervention that might work best.

So, by being well theorized and facilitated, it is entirely possible for community-based, context-adapted, flexible interventions to be evaluated usefully, even in randomized trials. This is a point that has been argued in theory[56] and recently demonstrated in practice.[59]

Myth 5. This Is Just Health Promotion Research With a Different Name

There is an extraordinary legacy of work in health promotion research that informs the way in which we should conceive of and measure the process and impact of interventions in population health.[60] However, population health intervention research is wider.

The difference between health promotion research and intervention research in population health, as PHIRIC has defined it, is the *intentionality* of the intervention. Health promotion research is focused on interventions designed to improve health. Intervention research in population health is the umbrella term that also includes explorations of the health effects of interventions in sectors outside of health designed for other purposes, such as increased transport usage. This is commonly known as a health impact assessment. An advantage in bringing a closer alliance of the two domains is that methods from one can inform the other. For example, mathematical modelling with large secondary data sets is a common means to explore the health effects of economic policy.[61] Such methods are less well known in mainstream health promotion journals but could be better used. By contrast, exposure measurement with regard to an intervention, and all the subtleties enshrined in the notion of intensity of "preventive dose", has been well developed in the health promotion literature.[62] However, it appears to be less well captured in fields outside of health promotion, where exposure may only be defined dichotomously (i.e., program deemed to be present or not).

Myth 6. Intervention Research Is Too Expensive

It is common to deplore the high costs of intervention trials, demonstration projects or participatory research and to plea for resources to be spent in other ways, but the truth is that we do not know whether intervention research is any more expensive than descriptive or analytic research in population health. Certainly, when an intervention study fails to record a reduction in a health problem, there always seems to be attention drawn to how much it cost to find this out. But it is unclear whether, if one counted up the costs of all the cross-sectional studies and cohort studies that have been chasing various risk factors over the years, those results are any less costly or of any more value. Overall, we do not have a

system for monitoring the added value of *any* particular study or field of study to insights in population health. Perhaps it is time that some metric was devised.

The promise ahead

This paper began with a reference to John Snow in the 1850s as an exemplar of a population health intervention researcher testing a theory. As it happened, his colleagues at the time thought that his theory of the source of the cholera was too exclusive, and it took another decade for more evidence to be gathered and for actions to be taken to control contaminated water sources.[63] We followed up with a reference to causal reasoning as it involved the decomposition of phenomena to mechanisms and sequences of events. That was an illustration of reasoning primarily attributed to the 16th century contributions of Renee Descartes. But by the mid-20th century, a new way of thinking called "systems thinking" emerged, emphasising the view of living organisms as integrated wholes whose properties cannot be reduced to those of the smaller parts.[64]

Our point is that intervention research has to extend itself to understanding and accelerating the uptake of new practices and recognizing that the reasoning processes we use in science are under constant interrogation. In intervention research, investigators are adopting system-thinking approaches,[65] and there are investigators putting "realist" views claiming that "theory-based" views can be contrasted with "scientific method", which is presumed to not be based on theory sufficiently. Agent-based ways of conceiving interventions are being contrasted with expert models. Views are hotly contested.[66]

In the meantime, the ever-growing burden of disease demands that we design effective interventions and put them into practice. Our plea is for systems to be created in Canada that attract the best minds and the best energies in the country to solving these issues. There has never been a more stimulating or crucial time to act. There is no point in changing thinking in population health, if we cannot change history with it.

REFERENCES

1. Hawe P, Shiell A. Using evidence to expose the unequal distribution of problems and the unequal distribution of solutions. *Eur J Public Health* 2007;17(5):413.
2. MacMahon B, Pugh TF. *Epidemiology: Principles and Methods*. Boston, MA: Little Brown, 1970.
3. Institute of Population and Public Health, Canadian Institutes of Health Research. Population Health Intervention Research Initiative for Canada ("PHIRIC") Workshop Report. Ottawa, ON: CIHR. Available online at: www.cihr-irsc.gc.ca/e/33515.html (Accessed December 2008).
4. Porter D. *Health, Civilisation and the State. A History of Public Health from Ancient to Modern Times*. London, UK: Routledge, 1999.
5. Fassin D. *L'espace politique de la santé. Essai de généalogie*. Paris: Presses Universitaires de France, 1996.
6. Wagenaar AC, Webster DW. Preventing injuries to children through compulsory automobile safety seat use. *Pediatrics* 1986;78:662-72.
7. Mills JL, Signore C. Neural tube defects rates before and after food fortification with folic acid. *Birth Defects Res* 2004;70(11):8445-55.
8. Mathee A, Rollin H, von Schirnding Y, Levin J, Naik I. Reductions in blood lead levels among school children following the introduction of unleaded petrol in South Africa. *Environ Res* 2006;100(3):319-22.
9. Institute of Population and Public Health, Canadian Institutes of Health Research. Mapping and Tapping the Wellsprings of Health. Strategic Plan 2002-2007. Ottawa: CIHR.
10. Rose G. *The Strategy of Preventive Medicine*. Oxford, UK: Oxford University Press, 1992.
11. Mischen PA, Sinclair TAP. Making implementation more democratic through action implementation research. *J Public Admin Res Theory* 2009;19:145-64.
12. O'Toole LJ. The theory-practice issue in policy implementation research. *Public Admin* 2004;82(2):309-29.
13. Noble CH. The eclectic roots of strategy implementation research. *J Business Res* 1999;45(2):119-34.
14. Nutbeam D, Bauman A. *Evaluation in a Nutshell*. Sydney: McGraw Hill, 2006.
15. Foy R, Eccles M, Grimshaw J. Why does primary care need more implementation research? *Fam Pract* 2001;18(4):353-55.
16. Hawe P, Degeling D, Hall J. *Evaluating Health Promotion. A Health Workers' Guide*. Sydney: Maclennan and Petty, 1990.
17. Brown CA, Milford RJ. The stepped wedge trial design: A systematic review. *BMC Med Res Methodol* 2006;6:54.
18. Suchman EA. *Evaluative Research*. New York: Russel Sage, 1967.
19. Weiss CH. *Evaluation,* 2nd ed. Upper Saddle River, NJ: Prentice Hall, 1998.
20. Zaza S, Briss PA, Harris KW. *The Guide to Community Preventive Services: What Works to Promote Health?* New York, NY: Oxford University Press, 2005.
21. Spinks A, Turner C, Nixon J, McLure R. 'WHO Safe Communities' model for the prevention of injury in whole populations. *Cochrane Database Syst Rev* Issue 2, Art. No. CD004445.
22. Sefton C. The NSW Safe Communities pilot projects – evaluation methodology. *N S W Public Health Bull* 2002;13(4):76-77.
23. Jackson N, Waters E and the Guidelines for Systematic Reviews in Health Promotion and Public Health Taskforce. The challenges of systematically reviewing public health interventions. *J Public Health Med* 2004;26:303-7.
24. Susser M. The tribulations of trials. Interventions in communities. *Am J Public Health* 1995;85:156-60.
25. Leventhal H, Safer MA, Cleary PD, Gutman M. Cardiovascular risk reduction by community based programs for lifestyle change: Comments on the Stanford study. *J Consult Clin Psychol* 1980;48:150-58.
26. Patton MQ. *Qualitative Research Methods*, 2nd ed. Newbury Park, CA: Sage, 1990.
27. Stokols D. Translating social ecological theory into guidelines for community health action. *Am J Health Promotion* 1996;10(4):282-98.
28. McLeroy K, Steckler A, Simons-Morton B, Goodman RM. Social science theory in health education: Time for a new model? *Health Educ Res* 1993;8(3):305-12.
29. Ingle MD, Klaus R. Competency based program evaluation: A contingency approach. *Evaluation and Program Planning* 1981;3:277-87.
30. Petticrew M, Whitehead M, Macintyre S, Graham H, Egan M. Evidence for public health policy on inequalities: 1. The reality according to policymakers. *J Epidemiol Community Health* 2004;58:811-16.
31. Patton MQ. *Utilisation-Focused Evaluation: The New Century Text*. Newbury Park: Sage, 1997.
32. Perry CL, Komro KA, Veblen-Mortenson S, Bosma LM, Farbakhsh K, Munson KA, et al. A randomised controlled trial of the middle and high school DARE and DARE plus programmes. *Arch Pediatr Adolesc Med* 2003;157:178-84.
33. Lord M. Truth or dare. A new drug course. *US News World Rep* 2001;130(8):30.
34. Birkeland S, Murphy-Graham E, Weiss C. Good reasons for ignoring good evaluation: The case of the drug abuse resistance education (DARE) program. *Evaluation and Program Planning* 2005;28(3):247-56.
35. Steckler A, Linnan L (Eds.). *Process Evaluation for Public Health Interventions and Research*. San Francisco, CA: Jossey Bass, 2002.
36. Goodman RM, Steckler AB. A model of institutionalisation of health promotion programs. *Fam Community Health* 1987;11:63-78.
37. Potvin L, Haddad S, Frohlich KL. Beyond process evaluation. In: Rootman I, Goodstadt M, Hyndman B, McQueen DV, Potvin L, Springett J, et al. (Eds.), *Evaluation in Health Promotion. Principles and Perspectives*. Copenhagen: WHO Regional Publications, European Series, 2001; No 92:45-62.
38. Bisset S, Cargo M, Delormier T, Macaulay AC, Potvin L. Legitimizing diabetes as a community health issue: A case analysis of an Aboriginal community in Canada. *Health Promotion Int* 2004;19:317-26.
39. Baker IR, Dennison BA, Boyer PS, Sellers KF, Russo TJ, Sherwood NA. An asset-based community initiative to reduce television viewing in New York State. *Prev Med* 2007;44:437-41.
40. Hawe P, Stickney EK. Developing the effectiveness of an intersectoral food policy coalition through formative evaluation. *Health Educ Res* 1997;12:213-25.
41. Levesque L, Guilbault G, Delormier T, Potvin L. Unpacking the black box: A deconstruction of the programming approach and physical activity intervention implemented in the Kahnawake Schools Diabetes Prevention Project. *Health Promot Pract* 2005;6:64-71.
42. Ho LS, Gittelsohn J, Harris SB, Ford E. Development of an integrated prevention program with First Nations in Canada. *Health Promot Int* 2006;21:88-97.
43. Corrigan M, Cupples ME, Smith SM, Byrne M, Leathem CS, Clerkin P, et al. The contribution of qualitative research to designing a complex intervention for secondary prevention of coronary heart disease in two different health care systems. *BMC Health Serv Res* 2006;6:90.
44. Wickizer TM, Wagner E, Cheadle A, Pearson D, Berry W, Maeser J, et al. Implementation of the Henry J. Kaiser Family Foundation's Community Health Promotion Grant Program: A process evaluation. *Milbank Q* 1998;77:121-47.
45. Cooke MB, Ford J, Levine J, Bourke C, Newell L, Lapidus G. The effects of city-wide implementation of "Second Step" on elementary school students' prosocial and aggressive behaviors. *J Primary Prev* 2007;28:93-115.
46. O'Loughlin JL, Paradis G, Gray-Donald K, Renaud L. The impact of a community-based heart disease prevention program in a low-income inner-city neighbourhood. *Am J Public Health* 1999;89:1819-26.

47. Wagenaar AC, Erickson DJ, Harwood EM, O'Malley PM. Effects of state coalitions to reduce underage drinking: A national evaluation. *Am J Prev Med* 2006;31:307-15.
48. Potvin L, Goldberg C. Deux rôles joués par l'évaluation dans la transformation de la pratique en promotion de la santé. In : O'Neill M, Dupéré S, Pederson A, Rootman I (Eds.), *La promotion de la santé au Canada et au Québec : Perspectives critiques*. Québec: Presses de l'Université Laval, 2006;457-73.
49. Minkler M. *Community Organising and Community Building for Health*. Rutgers University Press, 2004.
50. Riley T, Hawe P, Shiell A. Contested ground: How should qualitative evidence inform the conduct of a community intervention trial? *J Health Serv Res Policy* 2005;10(2):103-10.
51. Cook TD, Campbell DT. *Quasi Experimentation: Design and Analysis Issues for Field Settings*. Chicago, IL: Rand McNally, 1979.
52. Biglan A, Ary D, Wagenaar AC. The value of interrupted time series experiments for community intervention research. *Prev Sci* 2000;1(1):31-49.
53. Zerger SL, Irizarry R, Peng RD. On time series analysis of public health and biomedical data. *Annu Rev Public Health* 2006;27:57-79.
54. Cargo M, Mercer SL. The value and challenge of participatory research: Strengthening its practice. *Annu Rev Public Health* 2008;29:325-53.
55. Minkler M, Wallerstein N. *Community Based Participatory Research for Health*. San Francisco: Jossey Bass, 2003.
56. Hawe P, Shiell A, Riley T. Complex interventions: How far 'out of control' should a randomised controlled trial be? *Br Med J* 2004;328:1561-63.
57. Hawe P, McKenzie N, Scurry R. Randomised controlled trial of the use of modified postal reminder card on the uptake of measles vaccination. *Arch Dis Childhood* 1998;79:136-40.
58. Flay BR. Efficacy and effectiveness trials (and other phases of research) in the development of health promotion programs. *Prev Med* 1986;15(5):451-74.
59. Patton GC, Bond L, Carlin JB, Thomas L, Butler H, Glover S, et al. Promoting social inclusion in schools: A group-randomized trial of effects on student health risk behavior and well-being. *Am J Public Health* 2006;96(9):1582-87.
60. Green LW, Kreuter MW. *Health Promotion Planning: An Educational and Ecological Approach,* 3rd ed. Mountain View, CA: Mayfield, 1999.
61. Morrell S, Taylor R, Quine S, Kerr C. Suicide and unemployment in Australia 1907-1990. *Soc Sci Med* 1993;36(6):749-56.
62. Bartholomew LK, Parcel GS, Kok G, Gottlieb NH. *Planning Health Promotion Programs. An Intervention Mapping Approach*. San Francisco: Jossey Bass, 2006.
63. Harrison M. *Disease and the Modern World: 1500 to the Present Day*. Cambridge: Polity, 2004.
64. Miller JH, Page SE. *Complex Adaptive Systems*. New Jersey: Princeton University Press, 2007.
65. Best A, Moor G, Holmes B, Clark PI, Brice T, Lieschow S, et al. Health promotion dissemination and system thinking: Towards an integrative model. *Am J Health Behav* 2003;27(Suppl):S206-S216.
66. Cook TD. The false choice between theory-based evaluation and experimentation. In: Rogers PJ, Hasci TA, Petrosino A, Huebner TA (Eds.), Program Theory in Evaluation: Challenges and Opportunities. *New Directions for Evaluation* 2000;87:27-34.

RÉSUMÉ

Les interventions populationnelles de santé comprennent l'ensemble des actions qui visent à modifier la distribution des risques à la santé en ciblant les conditions sociales, économiques et environnementales qui façonnent la distribution des risques. Sous forme de programmes et politiques, ces interventions peuvent provenir du secteur de la santé mais sont aussi souvent pilotées par d'autres secteurs comme l'éducation, le logement ou l'emploi. La recherche sur les interventions de santé des populations poursuivent l'objectif de documenter la valeur et les effets de ces interventions, les processus par lesquels les changements opèrent et les conditions qui favorisent les effets. Dans le domaine de la recherche en santé, des distinctions inutiles entre la recherche et l'évaluation ont retardé le développement des connaissances sur l'intervention de santé des populations et mené à une mauvaise intégration des données de recherche pour soutenir la pratique et les décisions concernant les programmes et politiques de santé des populations. Cet article déboulonne donc certains mythes pernicieux concernant la recherche sur les interventions, notamment relativement aux coûts associés, à ses visées et à la croyance en un rôle nécessairement marginal des communautés concernées pour développer des interventions efficaces. Cet article retourne aussi comme arbitraire et injustifiée la distinction traditionnelle entre la recherche sur les interventions et la recherche évaluative. En fait cet article montre que la recherche sur les interventions a tout à gagner d'un rapprochement avec la recherche évaluative et d'une intégration des méthodes de recherche appliquée provenant d'une diversité de disciplines.

Mots clés : Évaluation; intervention pour la santé des populations; pratique fondée sur des données probantes; recherche sur les interventions; santé des populations

OPEN ACCESS Freely available online

PLOS | MEDICINE

Policy Forum

Europe Needs a Central, Transparent, and Evidence-Based Approval Process for Behavioural Prevention Interventions

Fabrizio Faggiano[1]*, Elias Allara[1,2], Fabrizia Giannotta[3,4], Roberta Molinar[1], Harry Sumnall[5], Reinout Wiers[6], Susan Michie[7], Linda Collins[8], Patricia Conrod[9,10]

1 Department of Translational Medicine, Università del Piemonte Orientale Amedeo Avogadro, Novara, Italy, **2** School of Public Health, University of Turin, Torino, Italy, **3** Child and Baby Lab, Department of Psychology, Uppsala University, Uppsala, Sweden, **4** Department of Public Health Sciences, Karolinska Institutet, Stockholm, Sweden, **5** Centre for Public Health, Liverpool John Moores University, Liverpool, United Kingdom, **6** Department of Psychology, University of Amsterdam, Amsterdam, Netherlands, **7** Department of Clinical, Educational and Health Psychology, University College London, London, United Kingdom, **8** The Methodology Center and Department of Human Development and Family Studies, Pennsylvania State University, State College, United States of America, **9** Addictions Department Institute of Psychiatry, King's College London, London, United Kingdom, **10** Department of Psychiatry, University of Montreal, Montreal, Québec, Canada

Introduction

Seven risk factors account for 56.1% of the attributable disability-adjusted life years (DALYs) in western Europe: dietary risks, smoking, high blood pressure, high body mass index (BMI), physical inactivity, excessive alcohol consumption, and high fasting plasma glucose [1]. Although such a figure reflects the predominant burden of noncommunicable diseases (NCDs) in high-income countries, it is becoming a priority also in middle-income and low-income countries [2]. By addressing such risk factors, prevention and health promotion can play a major role in reducing the burden of NCDs. The crucial function of prevention in tackling the NCDs epidemic is shared globally, as highlighted by the WHO programme "Gaining Health" and, more recently, by the United Nations High-Level Meeting on NCDs—in which prevention has been included among the five priority actions needed globally and nationally to respond to the NCDs epidemic [3,4].

What distinguishes prevention of NCDs from the more traditional prevention activities of communicable diseases is the aim to avoid or change health-compromising behaviours or to promote healthy behaviours. Prevention of NCDs includes individual and environmental intervention, e.g., family-based interventions tackling alcohol misuse, national policies prohibiting indoor smoking, school-based education to foster correct eating behaviour, walking groups for children or adults, taxation of tobacco or alcohol products, and policies to limit junk food in vending machines on school premises. This aim is particularly critical and requires much more caution than traditional prevention practices, for example, those intended to

Policy Forum articles provide a platform for health policy makers from around the world to discuss the challenges and opportunities in improving health care to their constituencies.

Summary Points

- Prevention interventions tackling health-compromising behaviours have the potential to play a major role in reducing the burden of noncommunicable diseases in Europe and other areas of the world. However, in Europe, no prior evaluation is required for the implementation of prevention interventions, thus leading to widespread dissemination of potentially ineffective or harmful interventions.
- A central, transparent, evidence-based, context-aware, and research-oriented approval process for behavioural prevention interventions is likely to foster the implementation and dissemination of effective interventions in Europe.
- Similarly to medicine approval systems, such a new approval process could be based on four consequential phases evaluating the effect of the following: single components (phase 1); combinations of components (phase 2); the final intervention—comprising only components found effective in the previous phases—via large, multicentre, randomized trials whenever possible (phase 3); and the long-term effects as well as the effects in different contexts (phase 4).
- Once phase 3 shows convincing results, the intervention would be approved for delivery to its target population.
- An approval process for behavioural prevention interventions is likely to lead to positive consequences both for practice, by strengthening the role and impact of prevention in times of limited economic resources, and for research, by promoting the robust evaluation of all promising prevention interventions.

Citation: Faggiano F, Allara E, Giannotta F, Molinar R, Sumnall H, et al. (2014) Europe Needs a Central, Transparent, and Evidence-Based Approval Process for Behavioural Prevention Interventions. PLoS Med 11(10): e1001740. doi:10.1371/journal.pmed.1001740

Published October 7, 2014

Funding: The research leading to these findings has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013), under Grant Agreement n. 266813-Addictions and Lifestyle in Contemporary Europe-Reframing Addictions Project (ALICE RAP). Participant organizations in ALICE RAP can be seen at http://www.alicerap.eu/about-alice-rap/partners.html. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: fabrizio.faggiano@med.unipmn.it

control environmental pollution and infectious diseases.

In Europe, a number of official documents, both at the regional and country level, promote an evidence-based approach to prevention [3,5,6]. However, there is no regulatory system for the implementation of behavioural prevention interventions. This contrasts with the situation in clinical medicine, in which there is a long-standing culture of using robust evidence to inform commissioning and clinical decisions. The European Medicines Agency (EMA) and European national authorities manage a well-established, although not perfect [7], system for the assessment of safety and effectiveness of drugs.

With this contribution, we aim to initiate a debate about the need for a unique European evaluation and approval system of prevention interventions for health-compromising behaviours.

Need for a Rigorous Evaluation of Behavioural Prevention Interventions

In Europe, interventions for preventing health-compromising behaviours can be implemented and disseminated without any preliminary authorisation, whatever setting (school, family, and community), professional, or type of method and technology involved. This is of concern for both ethical and economic reasons. First—and contrary to common belief—prevention interventions are not just harmless or ineffective in the worst-case scenario. They may also be harmful. Iatrogenic effects have been observed in interventions tackling risky behaviours such as physical inactivity, substance misuse, early sexual intercourse, and juvenile delinquency [8–12]. It is ethically unacceptable that a prevention intervention could significantly increase BMI, or tobacco or alcohol use, or frequency of cannabis use, or pregnancies and sexually transmitted diseases. Quoting the "father" of evidence-based medicine, David Sackett, "[…] the presumption that justifies the aggressive assertiveness with which we go after the unsuspecting healthy must be based on the highest level of randomized evidence that our preventive manoeuvre will, in fact, do more good than harm" [13].

Second, resources are likely to be wasted if evidence of effectiveness is missing or not sought. Cochrane reviews on the prevention of risky behaviours [14–16] show that effective interventions are in the minority of those evaluated by randomized studies. There is no reason to presume that non-evaluated interventions may be more effective than those that underwent rigorous evaluation. The resource allocation in the development and delivery of ineffective interventions is of particular concern in these times, given Europe's overstretched health systems.

Need for Improving the Analysis and Description of Mechanisms of Behavioural Prevention Interventions

Prevention interventions for health-compromising behaviours usually target psychological, social, and organisational factors hypothesised to mediate the associations between intervention and behavioural outcomes. Although theories should play a crucial role in the design and evaluation of prevention interventions, there is a lack of awareness and consensus as to which theories should be applied and what method should be used [17,18]. Many interventions are an amalgam of approaches and contents that do not explicitly draw on formal theories; others mention theory but are not truly theory-driven and do not always adhere consistently to a theory's tenets, being driven by implicit common sense models of behaviour [19,20].

Moreover, interventions are usually delivered as a complex combination of components ("active ingredients" targeting different mediators), both in terms of contents, activities, techniques, and modes of delivery. However, the interventions are usually poorly described. Less than 30% of reports of randomized studies present a detailed description of the intervention allowing accurate replication and implementation, and fewer include descriptions of mechanisms of action [19,21]. In addition, complex interventions, composed of several components, are usually evaluated together in randomized studies, which makes it difficult to disentangle the effect of a single ingredient on mediators and behavioural outcomes. The failure to conceptualize, define, and describe intervention components and mediators restricts the potential for evaluation to add evidence about effective interventions and mechanisms of action. In addition, the effects of context are rarely recognized, reported, or analysed.

Not knowing why, how, and where prevention interventions work limits knowledge about generalizability and optimization of interventions. It also increases the cost of implementation, as non-essential mediators might be inappropriately targeted and non-essential components may be inadvertently included.

If mediators of the target behaviour are identified, it is easier to design intervention components that are more likely to be effective. If the intervention components most strongly associated with effectiveness are known, more accessible, practical, and lower-cost, yet still effective, prevention interventions can be elaborated and disseminated. Moreover, they can be adapted to meet local needs and implemented in situations that are less ideal than research circumstances.

The complexity of behavioural prevention interventions, together with the lack of accurate reporting of mechanisms of action and analysis of the effects of components and their interactions, has serious consequences for prevention science: new interventions and evaluations occur in relative isolation, limiting the possibility of building an incremental technology of prevention [19].

Identifying and Selecting Evidence-Based Behavioural Prevention Interventions: Current Situation

Prevention guidelines are uncommon and usually of mixed quality, and no national or international systems exist for the regulation of effective interventions. Prevention professionals usually have to search and appraise the literature by themselves if they want to select evidence-based interventions to transfer into practice.

There are some local experiences of public registries of evidence-based interventions in some areas of prevention [22–24]. However, at least two reasons suggest that these registries may not be enough to guide practice. First, they have a weak level of global authority; thus, they cannot limit the proliferation of unevaluated or harmful interventions. Secondly, they present a great variability in the level of evidence required to define an intervention as effective, and in the way they help dissemination. This is obviously a potential cause of uncertainty for decision-makers and implementers. Table 1 compares the criteria for intervention classification adopted by seven registries considered by Gandhi et al. [22], to which we added two European resources, the European Monitoring Centre for Drugs and Drug Abuse (EMCDDA) 's Best Practice Portal and the Dutch Recognition System [23,24]. Although evidence of efficacy and quality of evaluation are considered by all registries, aspects such as quality of programme

Table 1. Criteria used in most widespread registries of evidence-based prevention interventions.

	Criteria							
List of registries	**Evidence of efficacy**	**Quality of evaluation**	**Quality of programme goals**	**Quality of programme rationale**	**Quality of programme content and appropriateness**	**Quality of programme implementation methods**	**Educational significance[a]**	**Usefulness and replicability**
Blueprints	Yes	Yes	Yes	N/A	N/A	N/A	N/A	Yes
Drug strategies: Making the Grade	Yes	Yes	Yes	N/A	N/A	N/A	N/A	N/A
ED List	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Maryland Report	Yes	Yes	N/A	N/A	N/A	N/A	N/A	N/A
NIDA Guide	Yes	Yes	N/A	N/A	N/A	N/A	N/A	N/A
SAMHSA National Registry of Evidence-Based Programs and Practices	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes
Youth Violence: A Report of the Surgeon General	Yes	Yes	Yes	N/A	N/A	N/A	N/A	Yes
EMCDDA Best Practice Portal	Yes	Yes	Yes	No	Yes	Yes	No	No
Dutch Recognition System	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes

This table is partly based on the work by Gandhi et al. [22], with the addition of the EMCDDA Best Practice Portal [23] and the Dutch Recognition System [24].

[a]The application describes how the intervention is integrated into schools' educational mission.

Abbreviations: EMCDDA, European Monitoring Centre for Drugs and Drug Abuse; NIDA, National Institute on Drug Abuse; SAMHSA, Substance Abuse and Mental Health Services Administration; ED, US Department of Education; N/A, not available.

doi:10.1371/journal.pmed.1001740.t001

contents, programme implementation methods, and programme replicability are considered only in four out of nine registries. In this panorama the attempts to define standards of quality of prevention intervention, for example, from the Society for Prevention Research (see at www.preventionresearch.org) and from the EMCDDA (see at www.emcdda.europa.eu), do not appear to have had any visible effects.

Existing Frameworks for the Regulation of Interventions

Improving the regulation of prevention interventions for health-compromising behaviours to ensure that effective interventions are implemented and disseminated is likely to be challenging.

In clinical practice, authorization agencies such as the United States Food and Drug Administration (FDA) and the European Medicine Agency (EMA) are appointed to manage an evidence-based evaluation process intended to guarantee that only safe and effective drugs will be approved for marketing. Although formal pathways are slightly different, for both agencies the process is based on a four-step evaluation: small trials to test pharmacodynamics, pharmacokinetics, and dosage (phase 1); medium trials for assessing efficacy and short-term effects (phase 2); large, randomized trials to evaluate effectiveness and side effects (phase 3); and post-marketing surveillance and additional studies for specific subgroups of patients and assessment of rare side effects (phase 4) [25]. Pharmaceutical companies apply for EMA and/or FDA approval by transmitting all the preclinical and clinical information obtained during the first three phases [26]. Approval is a necessary prerequisite for marketing a drug in Europe and in the United States.

A systematic approach to developing and evaluating complex prevention interventions, as the majority of prevention interventions are, has been developed by the United Kingdom's Medical Research Council (MRC) [27,28]. The first set of guidance proposed a process for the evaluation of complex interventions, which is logically consistent with the sequential phases of the drug approval process. The second set was based on a more sophisticated understanding of complexity and called for the defining of relevant intervention components, as well as underlying mechanisms and theories (modelling phase), testing acceptability and feasibility (pilot phase), evaluating effectiveness in an experimental study

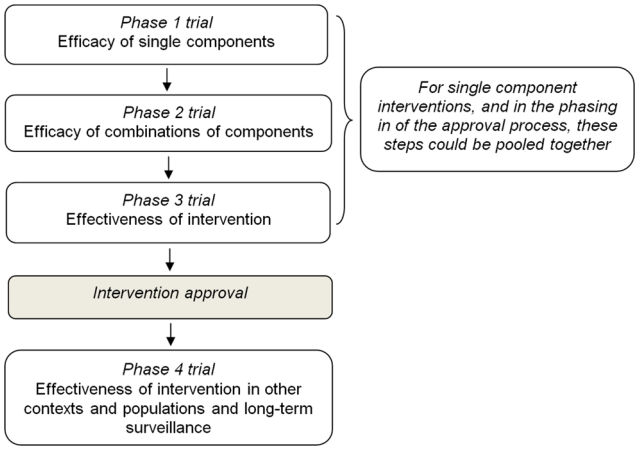


Figure 1. Proposal for a four-step evaluation and approval process of prevention interventions for health-compromising behaviours.
doi:10.1371/journal.pmed.1001740.g001

(evaluation phase), and assessing long-term effects in uncontrolled settings (implementation phase). The updated guidance in 2008 conceptualised the process as a cycle, emphasising the importance of considering implementation right at the beginning of the intervention development process.

Collins and colleagues addressed another issue that is not shared by drug registration systems: the complexity of interventions [29,30]. In order to evaluate the role of each single component of prevention interventions, they suggest adopting a multiphase optimization strategy, which may involve the application of a factorial design in order to assess the independent role of each component (see at http://methodology.psu.edu/ra/most).

These approaches, no matter how innovative, however, are not applicable to policy evaluation, a strong component of preventive strategies. In order to keep the same level of validity of the assessment, this requires tailored approaches to evaluation [31], as those developed for tobacco control [32].

A Proposal for a System of Evaluation and Approval of Behavioural Prevention Interventions

To tackle the overuse of interventions without scientific evidence and the underuse of effective interventions, Europe needs an approval system for prevention of health-compromising behaviours. This system would allow decision-makers and implementers to access the necessary information and materials to select the best prevention intervention for any specific need (e.g., target behaviour, population, setting, available resources, etc.). This system should be:

- Based on evidence. It should rely on the most valid evaluation approach for the specific intervention to assess. If randomized controlled trial would not be a feasible option, for example, for policy evaluation, the system should include alternative research designs that allow for relatively strong causal inferences (e.g., cohort design or interrupted time series design).
- Aware of context. Contextual moderators are of great importance for prevention of health-compromising behaviours and should be an essential part of the evaluation, as they may explain variations of effects across different contexts. They can help to describe *how* prevention interventions work and should be accurately identified and reported. Moreover, replications of evaluation studies in different contexts should be promoted and considered as an element of quality.
- Research-oriented. It should require an accurate reporting of underpinning theories, contents, mechanisms of action, and effects of single components on target behaviours to support the advancement of prevention science.
- Transparent and open access. All steps of evaluation should be transparently reported; descriptive information and complete data about evidence, benefits, risks, and variations related to different populations and contexts should be publicly available. The level of descriptive information must be sufficient to allow replications across different contexts with a high level of fidelity.
- Based on international cooperation. An international consensus on standards for releasing the certification of effectiveness is required to ensure widespread acceptance of this system in the scientific community. Therefore, a collaborative action of an extensive range of researchers, policy-makers, and health professionals is needed, as well as an extraordinary effort and mobilisation of resources.

In light of existing experiences, and taking into account the key characteristics described above, a four-phase evaluation and approval process could be proposed (Figure 1):

- Phase 1 should be aimed at evaluating the effect of single components on mediators and short-term outcomes through experimental or observational studies. This phase should also assess dosage features (e.g., delivery frequency, duration, etc.) and other delivery characteristics such as the appropriate age group.
- Phase 2 should be aimed at evaluating the effect of combinations of single components which passed phase 1 on short-term outcomes in the target population through a pilot experimental study.
- Phase 3 should be aimed at evaluating the effectiveness of the whole intervention, once individual components have shown evidence of effectiveness on short- and medium-term outcomes in phase 2. Whenever possible, an adequately powered, randomized, controlled design should be used to allocate individuals or target groups (e.g., schools, families) to study arms. But, since environmental interventions can hardly be evaluated by a randomized study, and they constitute a cornerstone of any comprehensive prevention strategy, such as smoking bans or taxation of sugar-sweetened beverage, they should be assessed with other studies of high validity, as for example, cohort studies or interrupted time series.

  Interventions found to be effective in this phase should be approved for implementation and dissemination.
- Phase 4 should be aimed at evaluating the effectiveness of approved interven-

tion in real-world settings (e.g., when delivered by a school team rather than a research team), the sustainability of effects on outcomes over a longer period of time and the long-term safety, and the replicability of effects on outcomes in different sociocultural contexts and populations, for which an adapted version of the intervention is usually needed.

Such a centralized system could be managed by a new public body, similar to the European Medicines Agency (EMA). Alternatively, an extended mandate to carry out such a process could be given to an existing and recognized public international agency or organization or to a network of research institutions coordinated by an international agency. The structural dimension of the proposed system cannot be easily estimated, but would be small. To make a rough estimation, an elaboration from the Cochrane Library can be of help: in 2012 the Library contained altogether 30 systematic reviews on primary prevention interventions; out of 503 interventions evaluated, only 171 (34.0%) showed at least one outcome favouring intervention [33]. Since Cochrane Library covers studies published in the last several decades and only studies showing positive results are expected to be submitted to such an approval process, these data suggest a few dozen interventions to be reviewed per year.

The funding requirements are a critical point: the amount depends too much on the ambition of the project, and cannot be estimated, even crudely. In any case, in analogy with a scientific journal, all the processes could be managed by a central editorial unit, supported by a network of referees, which would considerably contain costs.

Once an intervention has been approved, it should be included in a repository of effective interventions. The system would provide all needed materials and contacts with developers and trainers, together with the necessary information to select the intervention fitting the prevention needs (such as target behaviour, population, and setting), and contextual constraints (e.g., availability of human resources, time, and funding) of practitioners, decision-makers, and policymakers. The approval of a specific intervention can be nothing else than a strong recommendation to use the intervention. Nevertheless, with the progress of the project, and once the repository is populated sufficiently to be useful for all major conditions, we could expect that, at a country level, specific policies could be elaborated in order to promote the adoption of approved interventions.

Conclusions

Prevention research has made considerable methodological advances in the past decades. This is not reflected in a parallel improvement of practice, largely due to a lack of regulatory systems for transferring evidence into practice.

A possible exception is the Framework Convention on Tobacco Control (www.who.int/fctc), with which WHO produced a strong frame of effective actions for tobacco control. However, such a convention still remains an exception and can be hardly expected to be reproduced for other risk behaviours. The need to address the overall deficit in rigorous evaluation of prevention interventions for health-compromising behaviours is thus pressing in all other fields of prevention in Europe and beyond.

This paper aims to initiate a debate about how best to develop a central, transparent, public, and evidence-based system of evaluation and approval of prevention interventions for health-compromising behaviours in Europe. A four-phase approval process is outlined and is intended to foster further discussion.

This approval process would result in a repository of effective prevention interventions to be recommended to European Union member states for adoption, in order to base prevention strategies on scientific evaluations. Policy-makers and people working in the prevention field would find in this repository interventions and programmes to address the prevention needs of the target populations, together with all documents and materials useful to apply them. Furthermore, the repository would be even more useful for non-European and developing countries having similar health problems, for which the building of any systematic evaluation system for prevention is not foreseen for obvious economic reasons.

To steer the evaluation activities of prevention interventions in a transparent approval system would be a great progress not only for prevention practice but also for prevention science: the approval system would encourage evaluation, without which an intervention would not be included; would contribute to the standardization of evaluation methods; and would make available to the scientific community all the reports of the assessments, including component evaluation and mediation analysis. This could have the power to strongly improve research and give a contribution towards progressive learning on how prevention works.

Finally, the supply of effective and efficient behavioural interventions to prevention practice and policy making, in Europe but also elsewhere, would likely be a cost-effective initiative with a large expected impact on population health.

Author Contributions

Conceived and designed the experiments: FF EA RM. Wrote the first draft of the manuscript: FF EA RM. Wrote the paper: FF EA RM FG HS RW SM LC PC. ICMJE criteria for authorship read and met: FF EA RM FG HS RW SM LC PC. Agree with manuscript results and conclusions: FF EA RM FG HS RW SM LC PC.

References

1. Institute for Health Metrics and Evaluation (2013) GBD Compare. Available: http://viz.healthmetricsandevaluation.org/gbd-compare/. Accessed 4 March 2014.
2. Beaglehole R, Bonita R, Horton R, Adams C, Alleyne G, et al. (2011) Priority actions for the non-communicable disease crisis. Lancet 377: 1438–1447. Available: http://www.ncbi.nlm.nih.gov/pubmed/21474174. Accessed 12 March 2012.
3. WHO Europe (2006) Gaining Health The European Strategy for the Prevention and Control of Noncommunicable Diseases. Copenhagen, Denmark: WHO Regional Office for Europe.
4. Beaglehole R, Bonita R, Alleyne G, Horton R, Li L, et al. (2011) UN High-Level Meeting on Non-Communicable Diseases: addressing four questions. Lancet 378: 449–455. Available: http://www.ncbi.nlm.nih.gov/pubmed/21665266. Accessed 12 March 2012.
5. EMCDDA (2013) European Drug Report. Luxembourg: Publications Office of the European Union. Available: http://www.emcdda.europa.eu/attachements.cfm/att_213154_EN_TDAT13001ENN1.pdf. Accessed 4 March 2014.
6. WHO Regional Office for Europe (2012) Health 2020: a European policy framework supporting action across government and society for health and well-being. Copenhagen, Denmark: WHO Regional Office for Europe.
7. Wieseler B, McGauran N, Kaiser T (2012) We need access to data on all clinical trials, not just some. BMJ 345: e8001. Available: http://www.ncbi.nlm.nih.gov/pubmed/23183067. Accessed 26 June 2014.
8. Werch CE, Owen DM (2002) Iatrogenic effects of alcohol and drug prevention programs. J Stud Alcohol 63: 581–590. Available: http://www.ncbi.nlm.nih.gov/pubmed/12380855. Accessed 16 July 2012.
9. Kirby D, Korpi M, Barth RP, Cagampang HH (1997) The impact of the Postponing Sexual Involvement curriculum among youths in California. Fam Plann Perspect 29: 100–108. Avail-

able: http://www.ncbi.nlm.nih.gov/pubmed/9179578. Accessed 3 June 2012.
10. Petrosino A, Turpin Petrosino C, John B (2004) "Scared Straight" and Other Juvenile Awareness Programs for Preventing Juvenile Delinquency. Campbell Syst Rev 2004.2. doi:10.4073/csr.2004.2.
11. Sallis JF, McKenzie TL, Alcaraz JE, Kolody B, Hovell MF, et al. (1993) Project SPARK. Effects of physical education on adiposity in children. Ann N Y Acad Sci 699: 127–136. Available: http://www.ncbi.nlm.nih.gov/pubmed/8267303. Accessed 2 July 2012.
12. Sloboda Z, Stephens RC, Stephens PC, Grey SF, Teasdale B, et al. (2009) The Adolescent Substance Abuse Prevention Study: A randomized field trial of a universal substance abuse prevention program. Drug Alcohol Depend 102: 1–10. Available: http://www.ncbi.nlm.nih.gov/pubmed/19332365. Accessed 19 April 2012.
13. Sackett DL (2002) The arrogance of preventive medicine. CMAJ 167: 363–364. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=117852&tool=pmcentrez&rendertype=abstract. Accessed 13 June 2012.
14. Faggiano F, Vigna-Taglianti FD, Versino E, Zambon A, Borraccino A, et al. (2005) School-based prevention for illicit drugs' use. Cochrane Database Syst Rev: CD003020. Available: http://www.ncbi.nlm.nih.gov/pubmed/15846647. Accessed 11 November 2013.
15. Foxcroft DR, Tsertsvadze A (2011) Universal multi-component prevention programs for alcohol misuse in young people. Cochrane Database Syst Rev: CD009307. Available: http://www.ncbi.nlm.nih.gov/pubmed/21901732. Accessed 11 November 2013.
16. Waters E, de Silva-Sanigorski A, Hall BJ, Brown T, Campbell KJ, et al. (2011) Interventions for preventing obesity in children. Cochrane database Syst Rev: CD001871. Available: http://www.ncbi.nlm.nih.gov/pubmed/22161367. Accessed 11 November 2013.
17. Michie S, van Stralen MM, West R (2011) The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 6: 42. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3096582&tool=pmcentrez&rendertype=abstract. Accessed 11 November 2013.
18. Noar SM, Zimmerman RS (2005) Health Behavior Theory and cumulative knowledge regarding health behaviors: are we moving in the right direction? Health Educ Res 20: 275–290. Available: http://www.ncbi.nlm.nih.gov/pubmed/15632099. Accessed 26 July 2011.
19. Michie S, Fixsen D, Grimshaw JM, Eccles MP (2009) Specifying and reporting complex behaviour change interventions: the need for a scientific method. Implement Sci 4: 40. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2717906&tool=pmcentrez&rendertype=abstract. Accessed 11 November 2013.
20. Hansen WB, Dusenbury L, Bishop D, Derzon JH (2007) Substance abuse prevention program content: systematizing the classification of what programs target for change. Health Educ Res 22: 351–360. Available: http://www.ncbi.nlm.nih.gov/pubmed/16963725. Accessed 31 August 2011.
21. Riley BL, MacDonald J, Mansi O, Kothari A, Kurtz D, et al. (2008) Is reporting on interventions a weak link in understanding how and why they work? A preliminary exploration using community heart health exemplars. Implement Sci 3: 27. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2413262&tool=pmcentrez&rendertype=abstract. Accessed 11 November 2013.
22. Gandhi AG, Murphy-Graham E, Petrosino A, Chrismer SS, Weiss CH (2007) The devil is in the details: examining the evidence for "proven" school-based drug abuse prevention programs. Eval Rev 31: 43–74. Available: http://www.ncbi.nlm.nih.gov/pubmed/17259575. Accessed 26 June 2012.
23. Bo A, Allara E, Ferri M (2011) Scientific evidence and practice: bridging the gap. A European Monitoring Center for Drugs and Drug Addiction (EMCDDA) project to promote Best Practice in Drug Addiction field. 19th Cochrane Colloquium; 20 October 2011; Madrid, Spain. Available: http://2011.colloquium.cochrane.org/abstracts/a1o4-scientific-evidence-and-practice-bridging-gap-european-monitoring-center-drugs-and-dr. Accessed 8 September 2014.
24. Brug J, van Dale D, Lanting L, Kremers S, Veenhof C, et al. (2010) Towards evidence-based, quality-controlled health promotion: the Dutch recognition system for health promotion interventions. Health Educ Res 25: 1100–1106. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2974836&tool=pmcentrez&rendertype=abstract. Accessed 23 March 2012.
25. Pharmaceutical Research and Manufacturers of America (n.d.) Drug Discovery And Development. Available: http://www.phrma.org/sites/default/files/pdf/PhRMA%20Profile%202013.pdf. Accessed 4 August 2011.
26. Lipsky MS, Sharp LK (2001) From idea to market: the drug approval process. J Am Board Fam Pract 14: 362–367. Available: http://www.ncbi.nlm.nih.gov/pubmed/11572541. Accessed 21 July 2011.
27. Campbell M, Fitzpatrick R, Haines A, Kinmonth AL, Sandercock P, et al. (2000) Framework for design and evaluation of complex interventions to improve health. BMJ 321: 694–696. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1118564&tool=pmcentrez&rendertype=abstract. Accessed 17 July 2012.
28. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, et al. (2008) Developing and evaluating complex interventions: the new Medical Research Council guidance. BMJ 337: a1655–a1655. Available: http://www.bmj.com/cgi/doi/10.1136/bmj.a1655. Accessed 15 March 2012.
29. Collins LM, Murphy S a, Nair VN, Strecher VJ (2005) A strategy for optimizing and evaluating behavioral interventions. Ann Behav Med 30: 65–73. Available: http://www.ncbi.nlm.nih.gov/pubmed/16097907.
30. Collins LM, Baker TB, Mermelstein RJ, Piper ME, Jorenby DE, et al. (2011) The multiphase optimization strategy for engineering effective tobacco use interventions. Ann Behav Med 41: 208–226. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3053423&tool=pmcentrez&rendertype=abstract. Accessed 11 November 2013.
31. Henry Rossi P, Lipsey MW, Freeman HE (2004) Evaluation: a systematic approach. New York: SAGE.
32. International Agency for Research on Cancer (2008) IARC Handbooks of Cancer Prevention in Tobacco Control. Volume 12. Lyon. Available: http://www.iarc.fr/en/publications/pdfs-online/prev/handbook14/handbook14.pdf.
33. Diego Concina (2012) An overview of Cochrane Systematic Reviews on NCD prevention - preliminary results [Thesis]. Alessandria, Italy: Università del Piemonte Orientale Amedeo Avogadro.

Original Investigation

Youth Problem Behaviors 8 Years After Implementing the Communities That Care Prevention System
A Community-Randomized Trial

J. David Hawkins, PhD; Sabrina Oesterle, PhD; Eric C. Brown, PhD; Robert D. Abbott, PhD; Richard F. Catalano, PhD

IMPORTANCE Community-based efforts to prevent adolescent problem behaviors are essential to promote public health and achieve collective impact community wide.

OBJECTIVE To test whether the Communities That Care (CTC) prevention system reduced levels of risk and adolescent problem behaviors community wide 8 years after implementation of CTC.

DESIGN, SETTING, AND PARTICIPANTS A community-randomized trial was performed in 24 small towns in 7 states, matched within state, assigned randomly to a control or intervention group in 2003. All fifth-grade students attending public schools in study communities in 2003-2004 who received consent from their parents to participate (76.4% of the eligible population) were included. A panel of 4407 fifth graders was surveyed through 12th grade, with 92.5% of the sample participating at the last follow-up.

INTERVENTIONS A coalition of community stakeholders received training and technical assistance to install CTC, used epidemiologic data to identify elevated risk factors and depressed protective factors for adolescent problem behaviors in the community, and implemented tested and effective programs for youths aged 10 to 14 years as well as their families and schools to address their community's elevated risks.

MAIN OUTCOMES AND MEASURES Levels of targeted risk; sustained abstinence, and cumulative incidence by grade 12; and current prevalence of tobacco, alcohol, and other drug use, delinquency, and violence in 12th grade.

RESULTS By spring of 12th grade, students in CTC communities were more likely than students in control communities to have abstained from any drug use (adjusted risk ratio [ARR] = 1.32; 95% CI, 1.06-1.63), drinking alcohol (ARR = 1.31; 95% CI, 1.09-1.58), smoking cigarettes (ARR = 1.13; 95% CI, 1.01-1.27), and engaging in delinquency (ARR = 1.18; 95% CI, 1.03-1.36). They were also less likely to ever have committed a violent act (ARR = 0.86; 95% CI, 0.76-0.98). There were no significant differences by intervention group in targeted risks, the prevalence of past-month or past-year substance use, or past-year delinquency or violence.

CONCLUSIONS AND RELEVANCE Using the CTC system continued to prevent the initiation of adolescent problem behaviors through 12th grade, 8 years after implementation of CTC and 3 years after study-provided resources ended, but did not produce reductions in current levels of risk or current prevalence of problem behavior in 12th grade.

TRIAL REGISTRATION clinicaltrials.gov Identifier: NCT01088542

JAMA Pediatr. 2014;168(2):122-129. doi:10.1001/jamapediatrics.2013.4009
Published online December 9, 2013.

Supplemental content at jamapediatrics.com

Author Affiliations: Social Development Research Group, School of Social Work, University of Washington, Seattle (Hawkins, Oesterle, Brown, Catalano); Educational Psychology, University of Washington, Seattle (Abbott).

Corresponding Author: J. David Hawkins, PhD, Social Development Research Group, School of Social Work, University of Washington, 9725 Third Ave NE, Ste 401, Seattle, WA 98115 (jdh@uw.edu).

Community-based efforts to prevent substance use, delinquency, and violence are an essential component of promoting health during adolescence and later life.[1,2] Communities That Care (CTC) is a prevention system that activates a coalition of stakeholders to develop and implement a science-based approach to prevention in the community to achieve collective impact on youth development community wide.[3,4] The CTC prevention system seeks to achieve this goal by increasing the use of tested, effective preventive interventions that address risk factors for adolescent problem behaviors prioritized by the community. This is expected to produce community-wide reductions in targeted risk factors and, in turn, decreased adolescent substance use, delinquency, and violence.[3,5]

The CTC system is different from other efforts to mobilize communities for the prevention of adolescent problem behaviors (eg, the Midwestern Prevention Project,[6-8] Project Northland,[9] Communities Mobilizing for Change on Alcohol,[10] the Community Trials Intervention to Reduce High Risk Drinking,[11,12] and PROSPER[13]). It does not focus exclusively on the prevention of alcohol use but rather on reducing shared risk factors for multiple behavior problems. It does not prescribe specific programs but trains the local coalition to choose programs from a menu of tested programs that best address the community's unique profile of risk and protection. In contrast to PROSPER, CTC does not prescribe who leads the prevention efforts but encourages stakeholders from a variety of organizations in the community to take leadership.

Results from a community-randomized trial of CTC support the CTC theory,[3,5] including increased adoption of a science-based approach to prevention[14-16] and implementation of a greater number of tested and effective prevention programs.[17] The trial also found that CTC lowered targeted risks for problem behavior and reduced the incidence and prevalence of seventh- and eighth-grade delinquency and substance use in a panel of youths followed up since fifth grade, 3 and 4 years after initial implementation of CTC.[18,19] These reductions continued to be observed 2 years later in 10th grade, 6 years after initial implementation of CTC and 1 year after support for the implementation of CTC had ended in the trial.[20]

This study tested the enduring effects of CTC on risk exposure and youth problem behaviors in 12th grade, 3 years after study-provided resources ended and 8 years after initial implementation of CTC in the trial. Although most CTC coalitions continued during the unsupported period,[21,22] very few of them used tested and effective prevention programs targeting high school students. The primary outcomes expected to be affected by CTC and examined in this study are targeted risk factors, substance use, delinquency, and violence.[23]

Methods

The Community Youth Development Study (CYDS)[5] is a community-randomized trial of CTC. Twenty-four communities in Colorado, Illinois, Kansas, Maine, Oregon, Utah, and Washington were matched in pairs within state on population size, racial and ethnic diversity, economic indicators, and crime rates. One community from within each matched pair was assigned randomly by a coin toss to either the intervention (CTC) or control group.[5] The CYDS communities are small to moderate-sized incorporated towns with their own governmental, educational, and law enforcement structures, ranging from 1500 to 50 000 residents.

Beginning in summer 2003, intervention communities received CTC training over 6 to 12 months by certified trainers. The CTC coalition members were trained to use data from cross-sectional CTC Youth Surveys of public school students in the community to prioritize risk factors to be targeted by tested and effective preventive actions.[24,25] Although CTC is designed for children and youths ages 0 to 18 years, CYDS communities were asked to focus their prevention plans on programs for youths aged 10 to 14 years and their families and schools so that possible effects on drug use and delinquency could be observed within the initial 5-year study period. Starting with the 2004-2005 school year and annually thereafter, community coalitions implemented between 1 and 5 preventive programs to address their prioritized risk factors. These included universal school-, family-, and community-based programs and selective school and community programs targeted at youths at elevated risk. The CYDS staff provided technical assistance and support for preventive interventions throughout the 5-year efficacy trial but stopped after the fifth year of the study. Control communities received data from CTC Youth Surveys administered in their schools every 2 years but received no resources, training, or technical assistance from the study.

Sample and Data Collection

The University of Washington Human Subjects Review Committee approved the protocol. Data were from a longitudinal panel of public school students in the 24 CYDS communities followed up from grade 5 through grade 12 (N = 4407).[23] Students were surveyed annually (2004-2011), except in 11th grade when students were tracked but not surveyed. The sample is sex balanced (50% male). Twenty percent of students identified as Hispanic or Latino, 64% were non-Hispanic white, 3% were non-Hispanic African American, 5% were non-Hispanic Native American, 1% were non-Hispanic Asian American, and 6% were of other ethnicities. All students in fifth-grade classrooms during the 2003-2004 school year in the 24 CYDS communities were eligible for participation in the study. Recruitment continued in grade 6 to increase the overall participation rate. Parents of 4420 students provided written informed consent to their participation in the study (76.4% of the total eligible population; 76.1% in CTC communities and 76.7% in control communities). The first wave of data collection (fifth grade, 2004) was a preintervention baseline assessment. The seventh wave of data was collected in spring 2011 when panel students progressing normally were in grade 12. At this point, 10 of the original 12 CTC coalitions were still active but had not received any support from the study for 3 years.[21,22] Tested and effective programs that were still being implemented in CTC communities during this unsupported period continued to be aimed primarily at middle school–aged adolescents (grades 5-9). Only 4 CTC communities implemented 1 of 3 substance abuse prevention programs aimed at high school–aged youths

(Project Toward No Drug Abuse, Class Action, or Communities Mobilizing for Change on Alcohol) during this period. Therefore, few panel students were exposed during the high school years to tested and effective prevention programs selected through the CTC process.

The longitudinal panel consists of 4407 students who completed a wave 1 or wave 2 survey. Students in the longitudinal panel who remained in the intervention or control communities for at least 1 semester were tracked and surveyed, even if they left the community, moved schools, or dropped out.[23] Seven students were deceased by the 12th-grade data collection and 2 students were permanently excluded from the sample owing to disability that precluded them from filling out the survey, leaving an active, still-living sample of 4398 students. Of the still-living sample members, 4068 (92.5%) completed the survey in 12th grade (2236 [93.2%] in CTC communities and 1832 [91.6%] in control communities) (**Figure**).

Students completed the Youth Development Survey,[26] a self-administered paper-and-pencil questionnaire designed to be completed in a class period. In 12th grade, 25.5% of participants completed the survey online because they were no longer attending school. Identification numbers but no names or other identifying information were included on the surveys. Participants received a $10 incentive check after completing the survey.

Measures

Targeted Risk Factors

The CTC coalitions prioritized 2 to 5 risk factors that were elevated in their community based on anonymous cross-sectional surveys of all assenting sixth- and eighth-grade students in their community.[27,28] Data used for targeting decisions were different from those used in the present analysis to evaluate intervention effects on risk factors. The cohort of fifth graders followed up in the trial did not participate in the cross-sectional surveys.

A targeted risk factor score was calculated for panel students in CTC communities by averaging the community-specific set of targeted risk factors. Items composing each risk factor scale were standardized within each year, and each scale was then standardized across years to facilitate pre-post comparisons. Because control communities did not prioritize risk factors using the CTC process, the average risk factor score in control communities was calculated using the set of targeted risk factors identified in its matched CTC community. For example, for students in community pair A, the targeted risk factor score was the average of scale scores for family conflict, antisocial friends, peer rewards for antisocial behavior, attitudes favorable to antisocial behavior, and rebelliousness; for students in community pair B, the targeted risk factor score was calculated based on scale scores for low commitment to school, family conflict, and antisocial friends (eTable 1 in Supplement shows the community-specific sets of targeted risk factors for all intervention communities).

Substance Use

Students reported their lifetime and past-month use of substances in grades 5 through 12 and past-year substance use in grade 12. Based on these prospective data, we examined sustained abstinence from any substance use, use of gateway drugs (alcohol, cigarettes, or marijuana), and binge drinking (having ≥5 drinks in 1 occasion) through grade 12 to assess the overall effect of CTC on preventing substance use. Cumulative incidence was examined for substances where use by grade 12 was less common than nonuse (ie, ≥50% of the sample reported never using by grade 12). The 12th-grade prevalence rates in the past month and the past year were computed for individual substances as well as composite indices of any substance use and gateway drugs (alcohol, cigarettes, or marijuana).

Delinquent and Violent Behavior

Each year, students reported participation in 7 delinquent and violent acts (stealing, damaging property, shoplifting, attacking someone with intent to harm, carrying a handgun, being arrested, and beating up someone so badly that he or she probably needed medical attention). A subset of the delinquency items (attacking someone with intent to harm, carrying a handgun, and beating up someone) was used to measure violent behavior. We computed sustained abstinence from delinquency and cumulative incidence of violence through spring of grade 12 as well as the past-year prevalence of both outcomes in grade 12. We also examined the number of different delinquent acts (ranging from 0-7) and different violent behaviors (ranging from 0-3) in the past year in grade 12.

Student and Community Characteristics

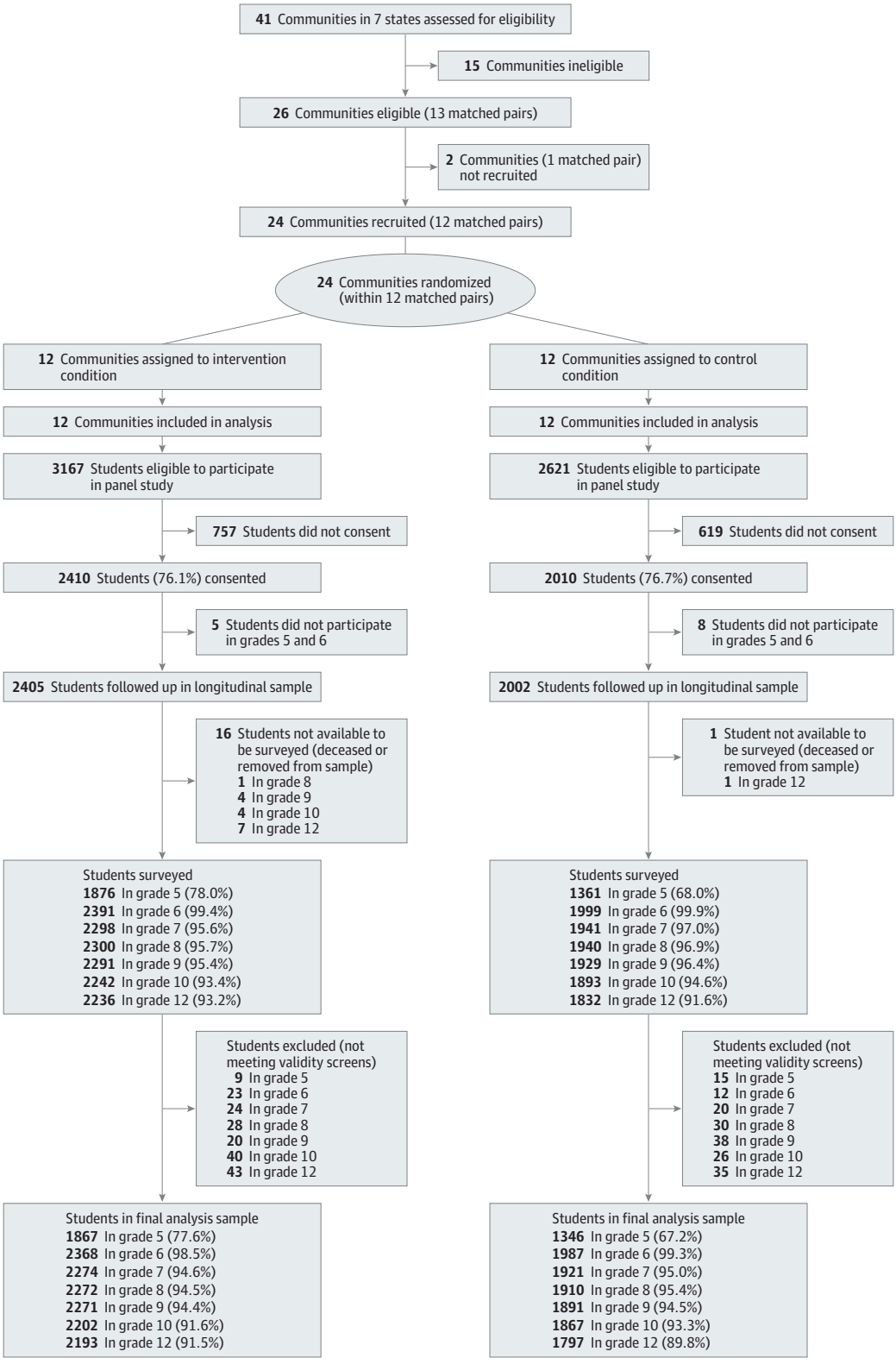
Student-level covariates included age, sex, race (white vs nonwhite), Hispanic ethnicity (Hispanic vs non-Hispanic), parental education, attendance at religious services during grade 5 (on a scale of 1 [never] to 4 [about once a week or more]), and rebelliousness in grade 5 (mean of 3 items; Cronbach $\alpha = .69$). Community-level covariates included the total population of students in the community (mean [SD], 2628 [1917]) and the percentage of students eligible for free or reduced-price school lunch (mean [SD], 38.2% [13.8%]).

Analysis Sample and Missing Data Procedures

Overall, 4021 students (91.4%) in the active, still-living sample participated in at least 6 of the 7 waves of data collection, and item nonresponse was small (<1%). Based on validity criteria (eg, reporting not being honest and using a fictitious drug), 78 students were excluded from the analysis sample in grade 12 (35 students in control and 43 students in CTC communities), resulting in valid data from 3990 students in 12th grade (90.7% of the active, still-living sample; 1797 students [89.8%] in control communities and 2193 students [91.5%] in CTC communities).

Of all the data points involved in the analysis (sample size times number of variables),[29] 11.8% were missing (10.8% in the CTC group and 13.0% in the control group). Missing data were imputed using multiple imputations to obtain unbiased estimates of model parameters and their standard errors, assuming that data are missing at random.[30] Using NORM version 2.03 software (Pennsylvania State University), 40 data sets including data from all 7 waves were imputed separately by intervention group.[31] Analyses were conducted within each imputed data set and then averaged using Rubin's rules.[32]

Figure. Flow of Communities and Participants in the Randomized Trial



Statistical Analysis

Because communities rather than students were randomized within matched community pairs, the effect of CTC was estimated as the mean difference between intervention groups in community-level sustained abstinence, cumulative incidence, prevalence, and means. Because community random-

Table 1. Sustained Abstinence From Substance Use and Delinquency Through Spring of Grade 12 Among Baseline Noninitiators Comparing CTC and Control Communities[a]

Substance Use or Delinquency	Noninitiators at Baseline in Grade 5		Cumulative Abstinence by Grade 12		
	CTC, %	Control, %	CTC, %	Control, %	ARR (95% CI)[b]
Any drugs	72.0	70.6	24.5	17.6	1.32 (1.06-1.63)[c]
Gateway drugs	76.8	73.9	29.4	21.0	1.31 (1.06-1.63)[c]
Alcohol	79.7	76.7	32.2	23.3	1.31 (1.09-1.58)[c]
Cigarettes	92.6	90.5	49.9	42.8	1.13 (1.01-1.27)[c]
Marijuana	99.6	99.3	52.6	48.2	1.07 (0.96-1.19)
Binge drinking	99.0	98.7	50.4	43.9	1.11 (0.97-1.28)
Delinquency	80.1	76.9	41.7	33.0	1.18 (1.03-1.36)[c]

Abbreviations: ARR, adjusted risk ratio; CTC, Communities That Care.

[a] All figures represent averages across 40 imputed data sets. There were no significant baseline differences by intervention group.

[b] For abstinence in the CTC vs control group, adjusted for student and community characteristics.

[c] Statistically significant at $P < .05$ (2-tailed).

ization does not guarantee equivalent student populations or that community pairs will remain similar over time, all analyses were adjusted for student and community characteristics and the respective preintervention baseline measure of the outcome to improve the precision of estimated intervention effects.[23,33,34] All covariates were grand-mean centered.

Sustained abstinence and cumulative incidence were assessed among students who had not yet engaged in the behavior at baseline (grade 5). Current prevalence in the 12th grade and targeted risk factor scores were examined in the full sample.

Generalized linear mixed models[35,36] with random effects for intercepts were used to model variability in outcomes across 4407 students, 24 communities, and 12 community pairs. Linear regression was used to estimate mean differences between CTC and control communities in average levels of targeted risk factors in grade 12, adjusting for baseline levels of targeted risk. Poisson regression with a log link and binomial error distribution was used to estimate adjusted risk ratios for sustained abstinence, cumulative incidence, and current prevalence.[37,38] Adjusted odds ratios estimated using logistic regression can be found in eTables 2, 3, and 4 in the Supplement.

The statistical significance of intervention effects was tested with 9 *df* (number of community-matched pairs [12] minus the number of community-level covariates [2], minus 1) and a type I error rate of .05 (2-tailed). All analyses were conducted using HLM 7 software (Scientific Software International), and population-average results are reported.[39]

Results

Baseline Intervention Group Equivalence

There were no statistically significant baseline differences by intervention group in levels of average targeted risk factors, the incidence and prevalence of substance use, delinquency, violence, or the mean number of delinquent and violent acts.[18,23] Accretion and attrition were similar in both intervention groups.

Targeted Risk

The adjusted mean difference between intervention groups in the targeted risk factor score in grade 12, adjusting for baseline levels of targeted risk and student and community characteristics, was not statistically significant (adjusted mean difference = 0.07; 95% CI, −0.03 to 0.18; $P = .16$).

Sustained Abstinence and Cumulative Incidence

Youths in CTC communities were significantly more likely than youths in control communities to have abstained from any substance use and the use of gateway drugs through the spring of 12th grade (**Table 1**). The proportion of 12th graders who had never used alcohol and who had never smoked cigarettes was significantly higher in CTC communities than in control communities, but there was no statistically significant difference by intervention group in sustained abstinence or in cumulative incidence of other substances (Table 1 and **Table 2**). Youths in CTC communities were also significantly more likely than youths in control communities to avoid ever engaging in delinquent (Table 1) or violent (Table 2) behavior through the spring of 12th grade.

Past-Month and Past-Year Prevalence

The proportion of students in control and CTC communities who used drugs in the past month or the past year did not differ significantly, with the exception of ecstasy use (**Table 3**). Students in CTC communities were almost twice as likely to use ecstasy in the past month as students in control communities. There were no significant differences by intervention group in past-year prevalence of delinquency and violence (Table 3) or the number of different delinquent behaviors (adjusted risk ratio = 1.03; 95% CI, 0.89-1.19; $P = .67$) and the number of different violent acts (adjusted risk ratio = 0.98; 95% CI, 0.78-1.22; $P = .81$).

Discussion

The results of this study indicate that 8 years after implementation of CTC in communities and 3 years after study-provided technical assistance and resources ended, CTC continued to prevent initiation of alcohol and tobacco use, delinquency, and violence through 12th grade in a panel of students followed up from grade 5. However, as implemented in this study, CTC did not produce reductions in levels of risk or the prevalence of current drug use, delinquent behavior, or violent behavior in grade 12.

Communities chosen for this randomized trial of CTC were towns of 50 000 or fewer residents and do not include urban

Table 2. Cumulative Incidence of Substance Use and Violence by Grade 12 Among Baseline Noninitiators Comparing CTC and Control Communities[a]

Substance Use or Violence	Noninitiators at Baseline in Grade 5		Cumulative Incidence by Grade 12		
	CTC, %	Control, %	CTC, %	Control, %	ARR (95% CI)[b]
Smokeless tobacco	98.1	97.2	31.6	34.6	0.97 (0.82-1.15)
Inhalants	91.5	91.3	29.3	31.9	0.93 (0.81-1.07)
Prescription drugs[c]	98.6	98.4	29.4	29.3	0.98 (0.85-1.13)
Ecstasy/MDMA[c]	98.6	98.4	13.5	12.0	1.18 (0.86-1.63)
Cocaine[c]	98.6	98.4	9.6	11.2	0.94 (0.73-1.21)
LSD[c]	98.6	98.4	11.7	10.6	1.15 (0.90-1.46)
Stimulants[c]	98.6	98.4	6.4	6.8	0.96 (0.68-1.36)
Other illegal drugs	98.6	98.4	25.3	25.4	1.07 (0.89-1.29)
Violence	92.2	88.9	34.4	41.1	0.86 (0.76-0.98)[d]

Abbreviations: ARR, adjusted risk ratio; CTC, Communities That Care; LSD, lysergic acid diethylamide; MDMA, 3,4-methylenedioxy-*N*-methylamphetamine.

[a] All figures represent averages across 40 imputed data sets. There were no significant baseline differences by intervention group.

[b] For incidence in the CTC vs control group, adjusted for student and community characteristics.

[c] At baseline (fifth grade), respondents were asked if they had used any other illegal drugs beyond marijuana and inhalants. They were not asked specifically about the use of prescription drugs, ecstasy, cocaine, LSD, and stimulants. Analyses of these specific drugs in 12th grade were conducted among baseline noninitiators of other illegal drugs.

[d] Statistically significant at $P < .05$ (2-tailed).

Table 3. Grade 12 Prevalence of Past-Month and Past-Year Substance Use, Delinquency, and Violence in CTC and Control Communities[a]

Substance Use, Delinquency, or Violence	%		
	CTC	Control	ARR (95% CI)[b]
Past mo			
Any drugs	46.6	48.4	1.01 (0.83-1.21)
Gateway drugs	45.3	46.3	1.01 (0.84-1.21)
Alcohol	35.7	36.1	1.04 (0.85-1.28)
Cigarettes	22.7	24.3	0.94 (0.76-1.15)
Marijuana	21.9	19.7	1.09 (0.93-1.28)
Smokeless tobacco	8.8	10.8	0.83 (0.66-1.06)
Inhalants	1.5	1.1	1.37 (0.73-2.57)
Prescription drugs	7.3	5.1	1.44 (0.98-2.12)
LSD	2.2	1.5	1.41 (0.81-2.45)
Cocaine	1.4	1.0	1.52 (0.77-2.99)
Stimulants	0.7	0.9	0.84 (0.37-1.89)
Ecstasy/MDMA	2.6	1.4	1.89 (1.09-3.27)[c]
Other illegal drugs	3.5	2.5	1.39 (0.90-2.15)
Past 2 wk			
Binge drinking	17.3	19.7	0.94 (0.72-1.23)
Past y			
Gateway drugs	60.7	65.3	0.97 (0.82-1.14)
Alcohol	55.6	59.2	0.99 (0.83-1.18)
Cigarettes	33.5	35.7	0.97 (0.82-1.15)
Marijuana	34.2	33.7	0.99 (0.87-1.12)
Delinquency	28.7	29.8	1.02 (0.90-1.17)
Violence	10.4	11.6	0.97 (0.77-1.21)

Abbreviations: ARR, adjusted risk ratio; CTC, Communities That Care; LSD, lysergic acid diethylamide; MDMA, 3,4-methylenedioxy-*N*-methylamphetamine.

[a] All figures represent averages across 40 imputed data sets. There were no significant baseline differences by intervention group.

[b] For prevalence in the CTC vs control group, adjusted for student and community characteristics.

[c] Statistically significant at $P < .05$ (2-tailed).

or suburban populations. Findings of this study may not generalize to larger communities. Another limitation is that the effect of CTC was evaluated in only 12 matched community pairs, which may have limited power to detect smaller intervention effects. However, the study detected substantively meaningful risk reductions or increases in abstinence between 12% and 32%. Youths in CTC communities were 32% more likely than youths in control communities to abstain from any drug use through 12th grade; they were 31% more likely to avoid ever using any of 3 gateway drugs (alcohol, cigarettes, or marijuana). They were 18% more likely to have avoided delinquent behavior and 14% less likely to have engaged in violence. Twelfth graders in CTC communities also had a 31% higher probability than students in control communities of hav-

ing never drank alcohol and were 12% more likely to have never smoked cigarettes. These effect sizes are similar to those found earlier when the panel was in 8th grade and when the benefit to cost ratio was estimated to be $5.30 per $1.00 invested in CTC based on the prevention of smoking and delinquency.[40]

Another possible threat to the internal validity of the study is that all analyses are based on self-report data, which carry the risk of social desirability bias or dishonesty. It is important to note that although this study was not a blinded trial, communities, not students, were randomized into intervention groups. It is highly unlikely that students in the longitudinal panel were aware of the intervention group to which their community belonged; thus, it is unlikely that there was differential self-report bias by intervention group that might account for any observed trial benefits. Further, we used validity checks to exclude a small number of students each year (<2% of the sample) deemed to have provided inaccurate reports of their behavior. This exclusion rate did not differ by intervention group. Additionally, the prevalence of substance use in this study is comparable to national data for the same cohort of 12th graders in the Monitoring the Future study.[41]

The enduring effects of CTC through 12th grade were observed with little preventive programming targeting the high school years. Because CTC communities were asked to focus their prevention plans on programs for youths in grades 5 through 9, and continued to do so following study support, few students in the longitudinal panel were exposed to tested and effective programs beyond ninth grade. It is noteworthy that initiation of alcohol use, tobacco use, delinquency, and violence in the panel was prevented through 12th grade in CTC communities.

Targeting preventive interventions during middle school, a developmentally sensitive time for drug use and delinquency initiation, appears to have prevented the onset of alcohol and tobacco use, delinquency, and violence in the panel through high school. However, the present findings suggest that continued preventive interventions during high school may be needed to lower the current prevalence of substance use, delinquency, and violence among those who have initiated these behaviors. This suggestion is consistent with results of the randomized trial of Project Northland, a school- and community-based approach to preventing adolescent alcohol use. Perry et al[9] found significant positive effects of Project Northland during the active intervention phase in middle school, but alcohol use grew faster among youths in intervention communities than in control communities in grades 9 and 10 when little programming took place. Positive effects in reducing alcohol use were found again, however, after preventive interventions were introduced in grades 11 and 12.

The higher prevalence of past-month use of ecstasy among 12th-grade students in CTC communities compared with control communities is the only significant negative effect associated with CTC observed in this panel.[18-20] This result should be interpreted with caution as the estimation of this intervention effect is based on small numbers of students reporting ecstasy use. In 11 of the 12 control communities and in 7 of the 12 CTC communities, no more than 4 students reported past-month ecstasy use in 12th grade. When community pairs were compared, the prevalence of past-month ecstasy use was higher in the CTC community than in the control community in 8 of 12 pairs and lower in the CTC community than in the control community in 4 pairs. In the absence of specific hypotheses or other evidence that would explain a negative intervention effect, it is unclear whether the higher prevalence of ecstasy use in grade 12 in CTC communities is an iatrogenic effect attributable to the intervention.

Conclusions

Sustained effects of CTC on preventing the initiation of alcohol use, tobacco use, delinquency, and violence through 12th grade are important. These effects were sustained with little preventive programming targeted at high school students during a period in which communities experienced economic stress likely to threaten prevention efforts.[42] Lack of a developmental focus on preventive intervention during the high school years may explain why CTC communities did not reduce current levels of targeted risk factors or the current prevalence of drug use, delinquency, or violence in the panel in grade 12. It is possible that communities using the CTC system could affect these behaviors if they expanded the use of tested and effective preventive interventions developmentally through the high school years, although research is needed to confirm this suggestion.

ARTICLE INFORMATION

Accepted for Publication: August 15, 2013.

Published Online: December 9, 2013. doi:10.1001/jamapediatrics.2013.4009.

Author Contributions: Dr Hawkins had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
Study concept and design: All authors.
Acquisition of data: Hawkins.
Analysis and interpretation of data: Oesterle, Brown, Abbott.
Drafting of the manuscript: Hawkins, Oesterle, Brown.
Critical revision of the manuscript for important intellectual content: Hawkins, Oesterle, Abbott, Catalano.
Statistical analysis: Oesterle, Brown, Abbott.
Obtained funding: Hawkins.
Study supervision: Hawkins.

Conflict of Interest Disclosures: Dr Catalano is a board member of Channing Bete Co, distributor of Supporting School Success and Guiding Good Choices. These programs were used in some communities in the study that produced the data set used in this article. No other disclosures were reported.

Funding/Support: This work was supported by research grant R01 DA015183-03 from the National Institute on Drug Abuse, with cofunding from the National Cancer Institute, the National Institute of Child Health and Human Development, the National Institute of Mental Health, the Center for Substance Abuse Prevention, and the National Institute on Alcohol Abuse and Alcoholism.

Role of the Sponsor: The funding organizations had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Disclaimer: The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

Additional Contributions: David M. Murray, PhD, Ohio State University, provided paid statistical consultation on this project, but the authors are responsible for all analyses and results. Tanya Williams provided editorial help. We acknowledge the contributions of the communities participating

in the Community Youth Development Study and the collaborating state offices of drug abuse prevention in Colorado, Illinois, Kansas, Maine, Oregon, Utah, and Washington.

REFERENCES

1. US Department of Health and Human Services. Healthy People 2020 Adolescent Health Objectives. http://www.healthypeople.gov/2020/topicsobjectives2020/overview.aspx?topicid=2. Accessed March 3, 2011.

2. US Department of Health and Human Services. *The Surgeon General's Call to Action to Prevent and Reduce Underage Drinking.* Washington, DC: Office of the Surgeon General; 2007.

3. Hawkins JD, Catalano RF. *Communities That Care: Action for Drug Abuse Prevention.* San Francisco, CA: Jossey-Bass; 1992.

4. Kania J, Kramer M. Collective impact. *Stanford Soc Innov Rev*. 2011;9(1):36-41.

5. Hawkins JD, Catalano RF, Arthur MW, et al. Testing Communities That Care: the rationale, design and behavioral baseline equivalence of the Community Youth Development Study. *Prev Sci*. 2008;9(3):178-190.

6. Pentz MA, Dwyer JH, MacKinnon DP, et al. A multicommunity trial for primary prevention of adolescent drug abuse. *JAMA*. 1989;261(22):3259-3266.

7. Pentz MA, Trebow EA, Hansen WB, MacKinnon DP. Effects of program implementation on adolescent drug use behavior: the Midwestern Prevention Project (MPP). *Eval Rev*. 1990;14(3):264-289.

8. Chou CP, Montgomery S, Pentz MA, et al. Effects of a community-based prevention program on decreasing drug use in high-risk adolescents. *Am J Public Health*. 1998;88(6):944-948.

9. Perry CL, Williams CL, Komro KA, et al. Project Northland: long-term outcomes of community action to reduce adolescent alcohol use. *Health Educ Res*. 2002;17(1):117-132.

10. Wagenaar AC, Gehan JP, Jones-Webb R, Toomey TL, Forster JL. Communities Mobilizing for Change on Alcohol: lessons and results from a 15-community randomized trial. *J Community Psychol*. 1999;27(3):315-326.

11. Grube JW. Preventing sales of alcohol to minors. *Addiction*. 1997;92(suppl 2):S251-S260.

12. Holder HD, Gruenewald PJ, Ponicki WR, et al. Effect of community-based interventions on high-risk drinking and alcohol-related injuries. *JAMA*. 2000;284(18):2341-2347.

13. Spoth R, Redmond C, Shin C, Greenberg M, Clair S, Feinberg M. Substance-use outcomes at 18 months past baseline: the PROSPER Community-University Partnership Trial. *Am J Prev Med*. 2007;32(5):395-402.

14. Brown EC, Hawkins JD, Arthur MW, Briney JS, Abbott RD. Effects of Communities That Care on prevention services systems: findings from the Community Youth Development Study at 1.5 years. *Prev Sci*. 2007;8(3):180-191.

15. Brown EC, Hawkins JD, Arthur MW, Briney JS, Fagan AA. Prevention service system transformation using Communities That Care. *J Community Psychol*. 2011;39(2):183-201.

16. Rhew IC, Brown EC, Hawkins JD, Briney JS. Sustained effects of the Communities That Care system on prevention service system transformation. *Am J Public Health*. 2013;103(3):529-535.

17. Fagan AA, Arthur MW, Hanson K, Briney JS, Hawkins JD. Effects of Communities That Care on the adoption and implementation fidelity of evidence-based prevention programs in communities. *Prev Sci*. 2011;12(3):223-234.

18. Hawkins JD, Brown EC, Oesterle S, Arthur MW, Abbott RD, Catalano RF. Early effects of Communities That Care on targeted risks and initiation of delinquent behavior and substance use. *J Adolesc Health*. 2008;43(1):15-22.

19. Hawkins JD, Oesterle S, Brown EC, et al. Results of a type 2 translational research trial to prevent adolescent drug use and delinquency: a test of Communities That Care. *Arch Pediatr Adolesc Med*. 2009;163(9):789-798.

20. Hawkins JD, Oesterle S, Brown EC, et al. Sustained decreases in risk exposure and youth problem behaviors after installation of the Communities That Care prevention system in a randomized trial. *Arch Pediatr Adolesc Med*. 2012;166(2):141-148.

21. Gloppen KM, Arthur MW, Hawkins JD, Shapiro VB. Sustainability of the Communities That Care prevention system by coalitions participating in the Community Youth Development Study. *J Adolesc Health*. 2012;51(3):259-264.

22. Arthur MW, Gloppen KM, Hawkins JD. Sustainability of the Communities That Care prevention system. Paper presented at: 20th Annual Meeting of the Society for Prevention Research; May 30, 2012; Washington, DC.

23. Brown EC, Graham JW, Hawkins JD, et al. Design and analysis of the Community Youth Development Study longitudinal cohort sample. *Eval Rev*. 2009;33(4):311-334.

24. Arthur MW, Hawkins JD, Pollard JA, Catalano RF, Baglioni AJ Jr. Measuring risk and protective factors for substance use, delinquency, and other adolescent problem behaviors: the Communities That Care Youth Survey. *Eval Rev*. 2002;26(6):575-601.

25. Glaser RR, Van Horn ML, Arthur MW, Hawkins JD, Catalano RF. Measurement properties of the Communities That Care® Youth Survey across demographic groups. *J Quant Criminol*. 2005;21(1):73-102.

26. Social Development Research Group. *Community Youth Development Study, Youth Development Survey, 2005-2001, Grades 5-12.* Seattle: Social Development Research Group, University of Washington; 2011.

27. Arthur MW, Briney JS, Hawkins JD, Abbott RD, Brooke-Weiss BL, Catalano RF. Measuring risk and protection in communities using the Communities That Care Youth Survey. *Eval Program Plann*. 2007;30(2):197-211.

28. Briney JS, Brown EC, Hawkins JD, Arthur MW. Predictive validity of established cut points for risk and protective factor scales from the Communities That Care Youth Survey. *J Prim Prev*. 2012;33(5-6):249-258.

29. Graham JW, Hofer SM. Multiple imputation in multivariate research. In: Little TD, Schnabel KU, Baumert J, eds. *Modeling Longitudinal and Multi-group Data: Practical Issues, Applied Approaches, and Specific Examples.* Hillsdale, NJ: Lawrence Erlbaum & Associates; 2000:201-218, 269-281.

30. Schafer JL, Graham JW. Missing data: our view of the state of the art. *Psychol Methods*. 2002;7(2):147-177.

31. Graham JW, Taylor BJ, Olchowski AE, Cumsille PE. Planned missing data designs in psychological research. *Psychol Methods*. 2006;11(4):323-343.

32. Rubin DB. *Multiple Imputation for Nonresponse in Surveys.* New York, NY: Wiley; 1987.

33. Murray DM. *Design and Analysis of Group-Randomized Trials.* New York, NY: Oxford University Press; 1998.

34. Schafer JL, Kang J. Average causal effects from nonrandomized studies. *Psychol Methods*. 2008;13(4):279-313.

35. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *J Am Stat Assoc*. 1993;88(421):9-25.

36. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986;73(1):13-22.

37. Cummings P. Methods for estimating adjusted risk ratios. *Stata J*. 2009;9(2):175-196.

38. Localio AR, Margolis DJ, Berlin JA. Relative risks and confidence intervals were easily computed indirectly from multivariable logistic regression. *J Clin Epidemiol*. 2007;60(9):874-882.

39. Zeger SL, Liang KY, Albert PS. Models for longitudinal data: a generalized estimating equation approach. *Biometrics*. 1988;44(4):1049-1060.

40. Kuklinski MR, Briney JS, Hawkins JD, Catalano RF. Cost-benefit analysis of Communities That Care outcomes at eighth grade. *Prev Sci*. 2012;13(2):150-161.

41. Johnston LD, O'Malley PM, Bachman JG, Schulenberg JE. *Monitoring the Future: National Results on Adolescent Drug Use: Overview of Key Findings, 2011.* Bethesda, MD: National Institute on Drug Abuse; 2012.

42. Kuklinski MR, Hawkins JD, Plotnick RD, Abbott RD, Reid CK. How has the economic downturn affected communities and implementation of science-based prevention in the randomized trial of Communities That Care? *Am J Community Psychol*. 2013;51(3-4):370-384.

A gradient of childhood self-control predicts health, wealth, and public safety

Terrie E. Moffitt[a,b], Louise Arseneault[b], Daniel Belsky[a], Nigel Dickson[c], Robert J. Hancox[c], HonaLee Harrington[a], Renate Houts[a], Richie Poulton[c], Brent W. Roberts[d], Stephen Ross[a], Malcolm R. Sears[e,f], W. Murray Thomson[g], and Avshalom Caspi[a,b,1]

[a]Departments of Psychology and Neuroscience and Psychiatry and Behavioral Sciences, and Institute for Genome Sciences and Policy, Duke University, Durham, NC 27705; [b]Social, Genetic, and Developmental Psychiatry Research Centre, Institute of Psychiatry, King's College London, London SE5 8AF, United Kingdom; [c]Dunedin Multidisciplinary Health and Development Research Unit, Department of Preventive and Social Medicine, School of Medicine, and [g]Department of Oral Sciences and Orthodontics, School of Dentistry, University of Otago, Dunedin, New Zealand; [d]Department of Psychology, University of Illinois, Urbana-Champaign, Champaign, IL 61820; [e]Department of Medicine, McMaster University, Hamilton, ON, L8S4L8 Canada; and [f]Firestone Institute for Respiratory Health, Hamilton, ON, Canada L8N 4A6

Edited by James J. Heckman, University of Chicago, Chicago, IL, and approved December 21, 2010 (received for review July 13, 2010)

Policy-makers are considering large-scale programs aimed at self-control to improve citizens' health and wealth and reduce crime. Experimental and economic studies suggest such programs could reap benefits. Yet, is self-control important for the health, wealth, and public safety of the population? Following a cohort of 1,000 children from birth to the age of 32 y, we show that childhood self-control predicts physical health, substance dependence, personal finances, and criminal offending outcomes, following a gradient of self-control. Effects of children's self-control could be disentangled from their intelligence and social class as well as from mistakes they made as adolescents. In another cohort of 500 sibling-pairs, the sibling with lower self-control had poorer outcomes, despite shared family background. Interventions addressing self-control might reduce a panoply of societal costs, save taxpayers money, and promote prosperity.

life course | longitudinal | public policy

The need to delay gratification, control impulses, and modulate emotional expression is the earliest and most ubiquitous demand that societies place on their children, and success at many life tasks depends critically on children's mastery of such self-control. We looked at the lives of 1,000 children. By the age of 10 y, many had mastered self-control but others were failing to achieve this skill. We followed them over 30 y and traced the consequences of their childhood self-control for their health, wealth, and criminal offending.

Interest in self-control unites all the social and behavioral sciences. Self-control is an umbrella construct that bridges concepts and measurements from different disciplines (e.g., impulsivity, conscientiousness, self-regulation, delay of gratification, inattention-hyperactivity, executive function, willpower, intertemporal choice). Neuroscientists study self-control as an executive function subserved by the brain's frontal cortex (1, 2) and have uncovered brain structures and systems involved when research participants exert self-control (3). Behavioral geneticists have shown that self-control is under both genetic and environmental influences (4) and are now searching for genes associated with self-control (5). Psychologists have described how young children develop self-control skills (6, 7) and have traced population patterns of stability and change in self-control across the life course (8). Health researchers report that self-control predicts early mortality (9); psychiatric disorders (10); and unhealthy behaviors, such as overeating, smoking, unsafe sex, drunk driving, and noncompliance with medical regimens (11). Sociologists find that low self-control predicts unemployment (12) and name self-control as a central causal variable in crime theory (13), providing evidence that low self-control characterizes lawbreakers (14, 15).

Economists are now drawing attention to individual differences in self-control as a key consideration for policy-makers who seek to enhance the physical and financial health of the population and reduce the crime rate (16, 17). The current emphasis on self-control skills of conscientiousness, self-discipline, and perseverance arises from the empirical observation that preschool programs that targeted poor children 50 y ago, although failing to achieve their stated goal of lasting improvement in children's intelligence quotient (IQ) scores, somehow produced byproduct reductions in teen pregnancy, school dropout, delinquency, and work absenteeism (18).* To the extent that self-control influences outcomes as disparate as health, wealth, and crime, enhancing it could have broad benefits. Given that self-control is malleable, it could be a prevention target, and the key policy question becomes when to intervene to achieve the best cost–benefit ratio, in childhood or in adolescence (19, 20)? Regardless of its malleability, however, if low self-control is influential, policy-makers might exploit this by enacting so-called "opt-out" schemes that tempt people to eat healthy food, save money, and obey laws by making these the default options that require no effortful self-control. If citizens were obliged to opt out of default health-enhancing programs or payroll-deduction retirement savings schemes, individuals with low self-control should tend to take the easy option and stay in programs, because opting out requires unappealing effort and planning (21, 22). Similarly, the idea behind the crime-reduction policy of "target hardening" is to discourage would-be offenders by making law-breaking require effortful planning (e.g., antitheft devices require more advance planning to steal a car).

In the context of this timely, ubiquitous, and intense policy interest in self-control, we report findings from the Dunedin Multidisciplinary Health and Development Study, a longitudinal study of a complete birth cohort of 1,037 children born in one city in a single year, whom we have followed from birth to the age of 32 y with 96% retention (Fig. 1 and *SI Appendix*). Our study design is observational and correlational; this is in contrast to experimental behavioral-economics studies that ascertain the association between performance on laboratory self-control tasks (e.g., delay of gratification, discounting, intertemporal choice tasks) and behavioral proxy measures of wealth, health, and crime. Such laboratory experiments yield compelling information about self-control, although economists have debated whether behavior in the laboratory faithfully represents real-world behavior (23). The naturalistic Dunedin study complements experimental research on self-control by providing badly needed information about how

Author contributions: T.E.M. and A.C. designed research; T.E.M., L.A., N.D., R.J.H., R.P., B.W.R., S.R., M.R.S., W.M.T., and A.C. performed research; T.E.M., D.B., H.L.H., R.H., and A.C. analyzed data; and T.E.M. and A.C. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

See Commentary on page 2639.

[1]To whom correspondence should be addressed. E-mail: avshalom.caspi@duke.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1010076108/-/DCSupplemental.

*Heckman JJ, Malofeeva L, Pinto R (2010) Economics of Crime Working Group, National Bureau of Economics Summer Institute, July 30, 2010, Cambridge, MA.

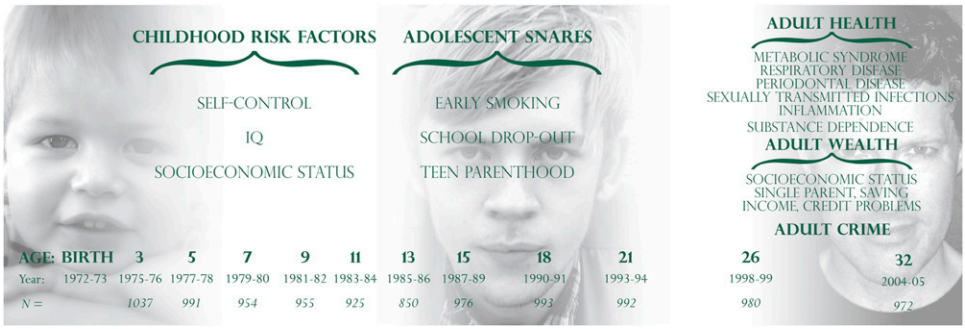


Fig. 1. Design of the Dunedin Multidisciplinary Health and Development Study.

well children's self-control, as it is distributed in the population, predicts real-world outcomes after children reach adulthood. We examined adult health outcomes, such as substance dependence, inflammation, and metabolic abnormalities (e.g., overweight, hypertension, cholesterol), because these are known early-warning signs for costly age-related diseases and premature mortality. We examined wealth outcomes, such as low income, single-parent child rearing, credit problems, and poor saving habits, because these are early warning signs for late-life poverty and financial dependence. We also examined convictions for crime, because crime control poses major costs to government.

The Dunedin study's birth-cohort members with low self-control and poor outcomes have not dropped out of the study. This enabled us to study the full range of self-control and to estimate effect sizes of associations for the general population, information that is requisite for informed policy making. The Dunedin study's design allowed us to address four policy-relevant hypotheses. First, we tested whether children's self-control predicted later health, wealth, and crime similarly at all points along the self-control gradient, from lowest to highest self-control. If self-control's effects follow a gradient, interventions that achieve even small improvements in self-control for individuals could shift the entire distribution of outcomes in a salutary direction and yield large improvements in health, wealth, and crime rate for a nation. Second, although this study did not include an intervention, some Dunedin study members moved up in the self-control rank over the years of the study, and we were able to test the hypothesis that improving self-control is associated with better health, wealth, and public safety. Third, because we assessed whether study members smoked tobacco as adolescents, left secondary school early, or became teen parents, we were able to test the hypothesis that children with low self-control make these mistakes as teenagers that close doors of opportunity and ensnare them in lifestyles harmful to their health and wealth as well as the public's safety. If self-control's influence is mediated through adolescents' mistakes, adolescence could be an ideal window for intervention policy. Fourth, because the Dunedin study assessed self-control as early as the age of 3 y, we were able to test the hypothesis that individual differences in preschoolers' self-control predict outcomes in adulthood. If so, early childhood would also be an intervention window.

Policy-making requires evidence that isolates self-control as the active ingredient affecting health, wealth, and crime, as opposed to other influences on children's futures, such as their intelligence or social class origins. Dunedin study data allowed the requisite statistical controls for IQ and social class. We also exploited another longitudinal study, a birth cohort of siblings, to ask whether the sibling in each pair who had lower self-control subsequently developed worse outcomes, despite both siblings having the same home and family. This design disentangles the individual child's self-control from all other features on which families differ (and which siblings share while growing up).

Results

This research aimed to ascertain whether childhood self-control predicts important adult outcomes along a population gradient. We assessed children's self-control during their first decade of life. Reports by researcher-observers, teachers, parents, and the children themselves gathered across the ages of 3, 5, 7, 9, and 11 y were combined into a single highly reliable composite measure. Mean levels of self-control were higher among girls than boys ($t = 8.39$, $P < 0.001$), but the health, wealth, and public safety implications of childhood self-control were equally evident and similar among boys and girls (*SI Appendix*, Table S1). We therefore combined the genders in all subsequent analyses (but controlled for gender). Dunedin study children with greater self-control were more likely to have been brought up in socioeconomically advantaged families ($r = 0.25$, $P < 0.001$) and had higher IQs ($r = 0.44$, $P < 0.001$), raising the possibility that low self-control could be a proxy for low social class origins or low intelligence. We thus tested whether childhood self-control predicted adults' health, wealth, and crime independent of their social class origins and IQ (the study design and measures are described in *SI Appendix*).

Predicting Health. When the children reached the age of 32 y, we assessed their cardiovascular, respiratory, dental, and sexual health as well as their inflammatory status by carrying out physical examinations and laboratory tests to assess metabolic abnormalities (including overweight), airflow limitation, periodontal disease, sexually transmitted infection, and C-reactive protein level, respectively. We summed these five clinical measures into a simple physical health index for each study member: 43% of study members had none of the biomarkers, 37% had one, and 20% had two or more. Childhood self-control predicted adult health problems (Table 1, model 1), even after accounting for social class origins and IQ (Table 1, model 2). *SI Appendix*, Table S1 shows associations between self-control and each individual health measure.

We also conducted clinical interviews with the study members at the age of 32 y to assess depression and substance dependence (tobacco, alcohol, and cannabis dependence as well as dependence on other street and prescription drugs), following the *Diagnostic and Statistical Manual of Mental Disorders*, 4th edition (DSM-IV) criteria (24). As adults, children with poor self-control were not at elevated risk for depression. They had elevated risk for substance dependence (Table 1, model 1), however, even after accounting for social class and IQ (Table 1, model 2). This longitudinal link between self-control and substance dependence was verified by people whom study members had nominated as informants who knew them well. As adults, children with poor self-control were rated by their informants as having alcohol and drug problems (Table 1).

Predicting Wealth. Childhood self-control also foreshadowed the study members' financial situations. Although the study members' social class of origin and IQ were strong predictors of their adult socioeconomic status and income, poor self-control offered significant incremental validity in predicting the socioeconomic position they achieved and the income they earned (Table 1). By the age of 32 y, 47% of study members had become parents. Childhood self-control predicted whether or not these study

Table 1. Does poor self-control in childhood lead to poor health, wealth-related problems, and criminal convictions in adulthood?

	Model 1: Baseline bivariate associations			Model 2: Co-occurring childhood risk factors hypothesis		
Adult outcomes and predictors	Coefficient	95% CI/SE	*P*	Coefficient	95% CI/SE	*P*
Health						
Physical health index*						
Low family SES	1.218	1.127–1.316	<0.001	1.154	1.058–1.258	0.001
Low IQ	1.224	1.133–1.323	<0.001	1.092	0.993–1.20	0.069
Low self-control	1.196	1.113–1.285	<0.001	1.111	1.020–1.209	0.016
Recurrent depression†						
Low family SES	1.038	0.876–1.229	0.667	0.955	0.790–1.153	0.629
Low IQ	1.232	1.031–1.470	0.022	1.208	0.978–1.493	0.080
Low self-control	1.187	0.944–1.419	0.059	1.099	0.849–1.352	0.369
Substance dependence index*						
Low family SES	1.343	1.184–1.523	<0.001	1.281	1.116–1.470	<0.001
Low IQ	1.218	1.074–1.382	0.002	1.012	0.870–1.177	0.880
Low self-control	1.299	1.156–1.460	<0.001	1.186	1.038–1.355	0.012
Informant-reported substance problems‡						
Low family SES	0.118	0.033	<0.001	0.076	0.036	0.033
Low IQ	0.081	0.034	0.014	−0.026	0.041	0.507
Low self-control	0.178	0.035	<0.001	0.169	0.039	<0.001
Wealth						
SES‡						
Low family SES	−0.266	0.033	<0.001	−0.124	0.034	<0.001
Low IQ	−0.400	0.033	<0.001	−0.312	0.039	<0.001
Low self-control	−0.263	0.035	<0.001	−0.082	0.038	0.023
Income‡						
Low family SES	−0.214	0.032	<0.001	−0.107	0.034	0.002
Low IQ	−0.291	0.033	<0.001	−0.199	0.039	<0.001
Low self-control	−0.238	0.034	<0.001	−0.112	0.038	0.002
Single-parent child rearing†§						
Low family SES	1.301	1.067–1.586	0.009	1.140	0.909–1.430	0.255
Low IQ	1.395	1.117–1.741	0.003	1.126	0.869–1.458	0.369
Low self-control	1.633	1.304–2.046	<0.001	1.479	1.147–1.908	0.003
Financial planfulness‡						
Low family SES	−0.151	0.032	<0.001	−0.090	0.036	0.011
Low IQ	−0.160	0.034	<0.001	−0.059	0.040	0.124
Low self-control	−0.195	0.034	<0.001	−0.141	0.039	<0.001
Financial struggles‡						
Low family SES	0.095	0.033	0.003	0.077	0.036	0.032
Low IQ	0.029	0.035	0.369	−0.068	0.041	0.078
Low self-control	0.152	0.035	<0.001	0.156	0.039	<0.001
Informant-reported financial problems‡						
Low family SES	0.131	0.033	<0.001	0.035	0.036	0.317
Low IQ	0.192	0.035	<0.001	0.077	0.041	0.045
Low self-control	0.274	0.034	<0.001	0.230	0.039	<0.001
Public safety						
Criminal conviction†						
Low family SES	1.578	1.337–1.863	<0.001	1.373	1.140–1.654	0.001
Low IQ	1.431	1.218–1.680	<0.001	0.967	0.792–1.179	0.737
Low self-control	1.830	1.559–2.148	<0.001	1.714	1.425–2.063	<0.001

Additional details are provided in *SI Appendix*, Table S1. SES, socioeconomic status.
*Incident-rate ratio.
†OR.
‡Standardized ordinary least squares regression coefficient.
§This analysis is restricted to 47% of the study members who have had a child.

members' offspring were being reared in one-parent vs. two-parent households (e.g., the study member was an absent father or single mother), also after accounting for social class and IQ (Table 1).

At the age of 32 y, children with poor self-control were less financially planful. Compared with other 32-y-olds, they were less likely to save and had acquired fewer financial building blocks for the future (e.g., home ownership, investment funds, retirement plans). Children with poor self-control were also struggling financially in adulthood. They reported more money-management difficulties and had accumulated more credit problems (Table 1). Poor self-control in childhood was a stronger predictor of these financial difficulties than study members' social class origins and IQ. This longitudinal link between self-control and self-reported financial problems was verified by informants who knew them well. As adults, children with poor self-control were rated by their informants as poor money managers (Table 1).

Predicting Crime. We obtained records of study members' court convictions at all courts in New Zealand and Australia by searching the central computer systems of the New Zealand Police; 24% of the study members had been convicted of a crime

by the age of 32 y. Children with poor self-control were more likely to be convicted of a criminal offense, even after accounting for social class origins and IQ (Table 1).

Self-Control Gradient. We observed a self-control gradient in which boys and girls with less self-control had worse health, less wealth, and more crime as adults than those with more self-control at every level of the distribution of self-control (Fig. 2). To document further that self-control relates to outcomes all along its gradient, we removed 61 study members who were diagnosed with attention-deficit hyperactivity disorder (a childhood psychiatric disorder of impaired impulse control) and repeated this analysis. The gradient associations in Fig. 2 remained unaltered. In addition, we tested whether self-control effects operate throughout the distribution or are confined to the least self-controlled children. We repeated analyses after removing children in the least self-controlled quintile and continued to observe significant linear associations. The self-control gradient was even apparent when we removed children in the least and most self-controlled quintiles (*SI Appendix*, Table S2).

An interesting question is what would happen if we were able to intervene and improve children's self-control. Would an increase in self-control predict better outcomes? Although the study did not include an experimental intervention, we were able to address this question by studying children who moved up the rank in their self-control from childhood to young adulthood. The childhood measure of self-control was significantly correlated with a personality measurement of self-control administered to our cohort in young adulthood ($r = 0.30$, $P \leq 0.001$), at a moderate magnitude, consistent with expectations (25) (*SI Appendix*). This stability coefficient implies that some children also changed their rank order in self-control. Moreover, those children who became more self-controlled from childhood to young adulthood had better outcomes by the age of 32 y, even after controlling for their initial levels of childhood self-control (*SI Appendix*, Table S3). As a caveat, it is not clear that natural history change of the sort we observed in our longitudinal study is equivalent to intervention-induced change. Nevertheless, these suggestive findings should stimulate consideration of interventions to raise self-control.

Self-Control and Adolescent Mistakes. Data collected at the ages of 13, 15, 18, and 21 y showed that children with poor self-control were more likely to make mistakes as adolescents, resulting in "snares"

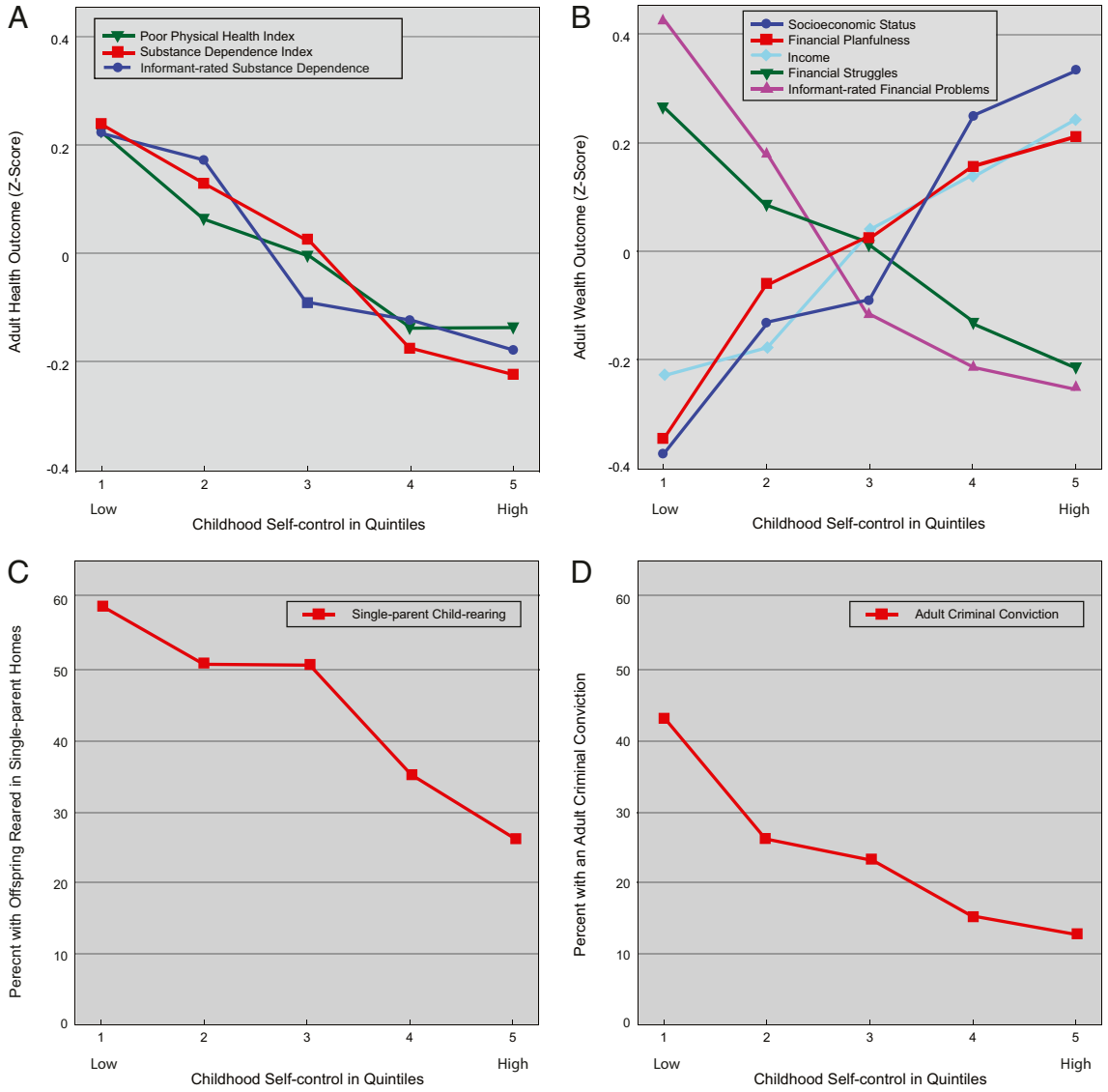


Fig. 2. Self-control gradient. Children with low self-control had poorer health (*A*), more wealth problems (*B*), more single-parent child rearing (*C*), and more criminal convictions (*D*) than those with high self-control.

that trapped them in harmful lifestyles. More children with low self-control began smoking by the age of 15 y [odds ratio (OR) = 1.69, 95% confidence interval (CI): 1.45–1.96], left school early with no educational qualifications (OR = 2.28, 95% CI: 1.92–2.72), and became unplanned teenaged parents (OR = 1.79, 95% CI: 1.40–2.29). The lower their self-control, the more of these snares they encountered (incident rate ratio = 1.48, 95% CI: 1.38–1.59) (*SI Appendix*, Fig. S1). In turn, the more snares they encountered, the more likely they were, as adults, to have poor health, less wealth, and criminal conviction (*SI Appendix*, Table S4). We tested whether snares explained the long-term predictive power of self-control in two ways. First, using statistical controls, we partialled out the portion of the association between childhood self-control and each adult outcome that was accounted for by adolescent snares. The snares attenuated the effect of self-control on health by 32%, substance dependence by 63%, socioeconomic status by 35%, income by 29%, single-parent child rearing by 61%, financial planfulness by 35%, financial struggles by 47%, and crime by 42%. The direct effect of self-control remained statistically significant for nearly every outcome measure, however (*SI Appendix*, Table S4). Second, we tested the association between childhood self-control and the adult outcomes among adolescents who did not encounter any snares, a so-called "utopian" control group (26). Again, prediction from childhood self-control to the adult measures remained significant even among this group of nonsmoking, non–teen-parent, high-school graduates (*SI Appendix*, Table S4).

How Early Can Self-Control Predict Health, Wealth, and Crime? Our composite measure of self-control in the Dunedin study included assessments from the age of 3–11 y. To answer this question, we isolated staff ratings of the children's self-control observed during 90-min data collection sessions at the research unit in the mid-1970s, when they were 3–5 y old (27). This standardized observational measure of preschoolers' self-control significantly predicted health, wealth, and convictions at the age of 32 y, albeit with modest effect sizes (*SI Appendix*, Table S5).

Sibling Comparisons. In the Dunedin study, statistical controls revealed that self-control had its own associations with outcomes, apart from childhood social class and IQ. Each Dunedin study member grew up in a different family, however, and their families varied widely on many features that affect children's outcomes. A compelling quasiexperimental research design that can isolate the influence of self-control is to track and compare siblings. Does the sibling with poorer self-control have worse outcomes than his or her more self-controlled sibling growing up in the same family environment? To apply this design, we turned to a second sample, the Environmental-Risk Longitudinal Twin Study (E-Risk), where we have been tracking a birth cohort of British twins since their birth in 1994 to 1995 with 96% retention (*SI Appendix*). When the E-Risk study twins were 5 y old, research staff rated each child on the same observational measure of self-control originally used with Dunedin study children as preschoolers. Although the E-Risk study children have been followed only up to age of 12 y, their self-control already forecasts many of the adult outcomes we saw in the Dunedin study. We applied sibling fixed-effects models to same-gender dizygotic pairs (n = 509 pairs) because they are no more alike than ordinary siblings (with the added advantages of being the same age and gender). Models showed that the 5-y-old sibling with poorer self-control was significantly more likely to begin smoking as a 12-y-old (a precursor of adult ill health; B = 0.07, SE = 0.003; $P < 0.03$), perform poorly in school (a precursor of adult wealth; B = −0.13, SE = 0.007; $P < 0.001$), and engage in antisocial behaviors (a precursor of adult crime; B = 0.09, SE = 0.007; $P = 0.007$), and these findings remained significant even after controlling for sibling differences in IQ (B = 0.07, SE = 0.003, $P = 0.02$ for smoking; B = −0.07, SE = 0.006, $P = 0.01$ for school performance; and B = 0.09, SE = 0.007, $P = 0.005$ for antisocial behavior).

Comment

Differences between individuals in self-control are present in early childhood and can predict multiple indicators of health, wealth, and crime across 3 decades of life in both genders. Furthermore, it was possible to disentangle the effects of children's self-control from effects of variation in the children's intelligence, social class, and home lives of their families, thereby singling out self-control as a clear target for intervention policy. Joining earlier longitudinal follow-ups (7, 9, 28), our findings imply that innovative policies that put self-control center stage might reduce a panoply of costs that now heavily burden citizens and governments.

Differences between children in self-control predicted their adult outcomes approximately as well as low intelligence and low social class origins, which are known to be extremely difficult to improve through intervention. Effects were marked at the extremes of the self-control gradient. For example, by adulthood, the highest and lowest fifths of the population on measured childhood self-control had respective rates of multiple health problems of 11% vs. 27%, rates of polysubstance dependence of 3% vs. 10%, rates of annual income under NZ $20,000 of 10% vs. 32%, rates of offspring reared in single-parent households of 26% vs. 58%, and crime conviction rates of 13% vs. 43%. This coincidence of low self-control with poor outcomes bolsters the rationale for opt-out programs by demonstrating that the segment of the adult population that is most inclined to avoid the effortful planning necessary to opt out of default programs (i.e., individuals with the lowest self-control) is the same segment of the adult population that accounts for excess costs to society in health care, financial dependency, and crime. Opt-out programs intended to exploit the laziness in all of us may work best for the least conscientious among us.

With respect to timing of programs to enhance self-control, our findings were consistent with "one-two punch" scheduling of interventions during both early childhood and adolescence (29). On the one hand, low self-control's capacity to predict health, wealth, and crime outcomes from childhood to adulthood was, in part, a function of mistakes our research participants made in the interim adolescent period. Adolescents with low self-control made mistakes, such as starting smoking, leaving high school, and having an unplanned baby, that could ensnare them in lifestyles with lasting ill effects. (Our choice of snares was not exhaustive, but we elected to study those that are already high-priority targets of adolescent education policy.) Thus, interventions in adolescence that prevent or ameliorate the consequences of teenagers' mistakes might go far to improve the health, wealth, and public safety of the population. On the other hand, that childhood self-control predicts adolescents' mistakes implies that early childhood intervention could prevent them. Moreover, even among teenagers who managed to finish high school as nonsmokers and nonparents, the level of personal self-control they had achieved as children still explained variation in their health, finances, and crime when they reached their thirties. Early childhood intervention that enhances self-control is likely to bring a greater return on investment than harm reduction programs targeting adolescents alone (30).

With respect to the scope of programs addressing self-control, our findings raise the question of whether early intervention to enhance self-control should take a targeted approach vs. a universal approach. Health, wealth, and crime outcomes followed a gradient across the full distribution of self-control in the population. If correct, the observed gradient implies room for better outcomes even among the segment of the population whose childhood self-control skills were somewhat above average. Universal interventions that benefit everyone often avoid stigmatizing anyone and also attract widespread citizen support. Testing this gradient in other population representative samples is a research priority. It has been shown that self-control can change (31). Programs to enhance children's self-control have been developed and positively evaluated, and the challenge remains to improve them and scale them up for universal dissemination (32–35). Understanding the key ingredients in self-control and how best to enhance them with a good cost–benefit ratio is a research priority.

Two cohorts born in different countries and different eras support the inference that individuals' self-control is a key ingredient in health, wealth, and public safety as well as a sensible policy target. That many Dunedin study members with low self-control had unplanned babies now growing up in low-income single-parent households reveals that one generation's low self-control disadvantages the next generation. Modern history is seeing a marked increase in human longevity, requiring individuals to pay more strategic attention to their health and wealth to avoid disability and poverty in old age (36). Modern history has also seen marked increases in food availability, sedentary occupations, access to harmful addictive substances, ease of divorce, self-management of retirement savings, and imprisonment of law-breakers. These historical shifts are enhancing the value of individual self-control in modern life, not just for well-being but for survival.

Methods

A more detailed report of the study designs, measures, and analyses is available in *SI Appendix*.

Dunedin Study Sample. Participants are members of the Dunedin Multidisciplinary Health and Development Study, which tracks the development of 1,037 individuals born in 1972–1973 in Dunedin, New Zealand.

Childhood Self-Control. Children's self-control during their first decade of life was measured using nine measures of self-control: observational ratings of children's lack of control (3 and 5 y of age) and parent, teacher, and self-reports of impulsive aggression, hyperactivity, lack of persistence, inattention, and impulsivity (5, 7, 9, and 11 y of age). The nine measures were positively and significantly correlated. Based on principal components analysis, the standardized measures were averaged into a single composite score (M = 0, SD = 1), comprising multiple ages and informants, with strong internal reliability $\alpha = 0.86$. *SI Appendix*, Table S6 shows that whether we examined self-control as measured by observers, teachers, parents, or children's self-reports, individual differences in childhood self-control were significantly related to each of the adult health, wealth, and public safety outcomes; that is, the results were not sensitive to the use of any particular source of information about children's self-control and were robust to the data source in measuring self-control.

Adult Outcomes. Health, wealth, and crime outcomes were assessed at age 32 y by physical examinations, blood tests, personal interviews, record searches, and informant reports.

Sample for Sibling-Comparison Analysis. Participants are members of the E-Risk study, which tracks the development of a nationally representative birth cohort of 2,232 twin children born in England and Wales in 1994–1995.

Childhood Self-Control at the Age of 5 Y. After completing the home visit when siblings were 5 y of age, examiners rated each twin on the measure of self-control that was originally used in the Dunedin study when the children in that study were 3 and 5 y of age (27). In this assessment procedure, the examiners evaluated the following behaviors: lability, low frustration tolerance, hostility, roughness, resistance, restlessness, impulsivity, fleeting attention, and lacking persistence. Each behavioral characteristic was defined in explicit terms, and the examiner evaluated whether each characteristic was not at all (0), somewhat (1), or definitely characteristic (2) of the child. The (interrater) reliability was 0.79.

Children's Outcomes at the Age of 12 Y. Children reported about their delinquent behavior and smoking. Children's educational performance was evaluated by their teachers, who rated each child's performance in English and mathematics.

ACKNOWLEDGMENTS. We thank the New Zealand Police. This research received support from the US National Institute on Aging Grants AG032282 and 2AG21178, National Institute of Mental Health Grant MH077874, National Institute of Child Health and Human Development Grant HD061298, National Institute of Dental and Craniofacial Research Grant DE015260, National Institute on Drug Abuse Grant DA023026, UK Medical Research Council Grants G0100527 and G0601483, and Economic and Social Research Council Grant RES-177-25-0013, as well as from the New Zealand Health Research Council, Lady Davis Fellowship of the Hebrew University, and Jacobs Foundation. L.A. has a Career Scientist Award from the UK Department of Health. A.C. is a Royal Society Wolfson Merit Award holder.

1. Eslinger PJ, Flaherty-Craig CV, Benton AL (2004) Developmental outcomes after early prefrontal cortex damage. *Brain Cogn* 55:84–103.
2. Stuss DT, Benson DF (1986) *The Frontal Lobes* (Raven, New York).
3. Hare TA, Camerer CF, Rangel A (2009) Self-control in decision-making involves modulation of the vmPFC valuation system. *Science* 324:646–648.
4. Bouchard TJ (2004) Genetic influence on human psychological traits. *Curr Dir Psychol Sci* 13:148–151.
5. Ebstein RP (2006) The molecular genetic architecture of human personality: Beyond self-report questionnaires. *Mol Psychiatry* 11:427–445.
6. Kochanska G, Coy KC, Murray KT (2001) The development of self-regulation in the first four years of life. *Child Dev* 72:1091–1111.
7. Mischel W, Shoda Y, Rodriguez MI (1989) Delay of gratification in children. *Science* 244:933–938.
8. Jackson JJ, et al. (2009) Not all conscientiousness scales change alike: A multimethod, multisample study of age differences in the facets of conscientiousness. *J Pers Soc Psychol* 96:446–459.
9. Kern ML, Friedman HS (2008) Do conscientious individuals live longer? A quantitative review. *Health Psychol* 27:505–512.
10. Caspi A, Moffitt TE, Newman DL, Silva PA (1996) Behavioral observations at age 3 years predict adult psychiatric disorders. Longitudinal evidence from a birth cohort. *Arch Gen Psychiatry* 53:1033–1039.
11. Bogg T, Roberts BW (2004) Conscientiousness and health behaviors: A meta-analysis. *Psychol Bull* 130:887–919.
12. Caspi A, Wright BRE, Moffitt TE, Silva PA (1998) Early failure in the labor market: Childhood and adolescent predictors of unemployment in the transition to adulthood. *Am Sociol Rev* 63:424–451.
13. Gottfredson M, Hirschi T (1990) *A General Theory of Crime* (Stanford Univ Press, Palo Alto, CA).
14. Caspi A, et al. (1994) Are some people crime-prone: Replications of the personality-crime relationship across countries, genders, races, and methods. *Criminology* 32:163–195.
15. White JL, et al. (1994) Measuring impulsivity and examining its relationship to delinquency. *J Abnorm Psychol* 103:192–205.
16. Heckman JJ (2007) The economics, technology, and neuroscience of human capability formation. *Proc Natl Acad Sci USA* 104:13250–13255.
17. Heckman JJ, Stixrud J, Urzua S (2006) The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *J Labor Econ* 24:411–482.
18. Heckman JJ (2006) Skill formation and the economics of investing in disadvantaged children. *Science* 312:1900–1902.
19. Carneiro P, Heckman JJ (2003) Human capital policy. *Inequality in America: What Role for Human Capital Policy?* eds Heckman JJ, Krueger A (MIT Press, Cambridge, MA).
20. Doyle O, Harmon CP, Heckman JJ, Tremblay RE (2009) Investing in early human development: Timing and economic efficiency. *Econ Hum Biol* 7:1–6.
21. Johnson EJ, Steffel M, Goldstein DG (2005) Making better decisions: From measuring to constructing preferences. *Health Psychol* 24(Suppl):S17–S22.
22. Thaler RH, Sunstein CR (2008) *Nudge: Improving Decisions About Health, Wealth, and Happiness* (Penguin Group, New York).
23. Falk A, Heckman JJ (2009) Lab experiments are a major source of knowledge in the social sciences. *Science* 326:535–538.
24. American Psychiatric Association (1994) *Diagnostic and Statistical Manual of Mental Disorders* (American Psychiatric Association, Washington, DC), 4th Ed.
25. Fraley RC, Roberts BW (2005) Patterns of continuity: A dynamic model for conceptualizing the stability of individual differences in psychological constructs across the life course. *Psychol Rev* 112:60–74.
26. Lubinski D (2009) Cognitive epidemiology: With emphasis on untangling cognitive ability and socioeconomic status. *Intelligence* 37:625–633.
27. Caspi A, Henry B, McGee RO, Moffitt TE, Silva PA (1995) Temperamental origins of child and adolescent behavior problems: From age three to age fifteen. *Child Dev* 66:55–68.
28. Deary IJ, Batty GD, Pattie A, Gale CR (2008) More intelligent, more dependable children live longer: A 55-year longitudinal study of a representative sample of the Scottish nation. *Psychol Sci* 19:874–880.
29. Knudsen EI, Heckman JJ, Cameron JL, Shonkoff JP (2006) Economic, neurobiological, and behavioral perspectives on building America's future workforce. *Proc Natl Acad Sci USA* 103:10155–10162.
30. Heckman J (2009) Stimulating the young. The American, http://www.american.com/archive/2009/august/stimulating-the-young. Accessed June 8, 2010.
31. Roberts BW, Walton KE, Viechtbauer W (2006) Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychol Bull* 132:1–25.
32. Greenberg MT (2006) Promoting resilience in children and youth: Preventive interventions and their interface with neuroscience. *Ann NY Acad Sci* 1094:139–150.
33. Layard R, Dunn J (2009) *A Good Childhood: Searching for Values in a Competitive Age* (Penguin, London).
34. National Scientific Council on the Developing Child (2007) *The Science of Early Childhood Development* (Harvard University Center on the Developing Child, Cambridge, MA).
35. Piquero AR, Jennings WG, Farrington DP (2010) On the malleability of self-control: Theoretical and policy implications regarding a general theory of crime. *Justice Q* 27:803–834.
36. Oeppen J, Vaupel JW (2002) Demography. Broken limits to life expectancy. *Science* 296:1029–1031.

BMC Medical Research Methodology

COMMENTARY Open Access

A tutorial on pilot studies: the what, why and how

Lehana Thabane[1,2*], Jinhui Ma[1,2], Rong Chu[1,2], Ji Cheng[1,2], Afisi Ismaila[1,3], Lorena P Rios[1,2], Reid Robson[3], Marroon Thabane[1,4], Lora Giangregorio[5], Charles H Goldsmith[1,2]

Abstract

Pilot studies for phase III trials - which are comparative randomized trials designed to provide preliminary evidence on the clinical efficacy of a drug or intervention - are routinely performed in many clinical areas. Also commonly know as "feasibility" or "vanguard" studies, they are designed to assess the safety of treatment or interventions; to assess recruitment potential; to assess the feasibility of international collaboration or coordination for multicentre trials; to increase clinical experience with the study medication or intervention for the phase III trials. They are the best way to assess feasibility of a large, expensive full-scale study, and in fact are an almost essential pre-requisite. Conducting a pilot prior to the main study can enhance the likelihood of success of the main study and potentially help to avoid doomed main studies. The objective of this paper is to provide a detailed examination of the key aspects of pilot studies for phase III trials including: 1) the general reasons for conducting a pilot study; 2) the relationships between pilot studies, proof-of-concept studies, and adaptive designs; 3) the challenges of and misconceptions about pilot studies; 4) the criteria for evaluating the success of a pilot study; 5) frequently asked questions about pilot studies; 7) some ethical aspects related to pilot studies; and 8) some suggestions on how to report the results of pilot investigations using the CONSORT format.

1. Introduction

The Concise Oxford Thesaurus [1] defines a *pilot project or study* as an *experimental, exploratory, test, preliminary, trial or try out* investigation. Epidemiology and statistics dictionaries provide similar definitions of a pilot study as a small scale

- " ***...test of the methods and procedures to be used on a larger scale*** if the pilot study demonstrates that the methods and procedures can work" [2];
- "...investigation designed to ***test the feasibility of methods and procedures*** for later use on a large scale or ***to search for possible effects and associations*** that may be worth following up in a subsequent larger study" [3].

Table 1 provides a summary of definitions found on the Internet. A closer look at these definitions reveals that they are similar to the ones above in that a pilot study is synonymous with a feasibility study intended to guide the planning of a large-scale investigation. Pilot studies are sometimes referred to as "vanguard trials" (i.e. pre-studies) intended to assess the safety of treatment or interventions; to assess recruitment potential; to assess the feasibility of international collaboration or coordination for multicentre trials; to evaluate surrogate marker data in diverse patient cohorts; to increase clinical experience with the study medication or intervention, and identify the optimal dose of treatments for the phase III trials [4]. As suggested by an African proverb from the Ashanti people in Ghana "*You never test the depth of a river with both feet*", the main goal of pilot studies is to assess feasibility so as to avoid potentially disastrous consequences of embarking on a large study - which could potentially "drown" the whole research effort.

Feasibility studies are routinely performed in many clinical areas. It is fair to say that every major clinical trial had to start with some piloting or a small scale investigation to assess the feasibility of conducting a larger scale study: critical care [5], diabetes management

* Correspondence: thabanl@mcmaster.ca
[1]Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton ON, Canada

Table 1 Some Adapted Definitions of Pilot Studies on the Web (Date of last access: December 22, 2009)

Definition*	Source
A trial study carried out before a research design is finalised **to assist in defining the research question** or **to test the feasibility, reliability and validity of the proposed study design**	http://www.cirem.org.uk/definitions.html
A smaller version of a study is carried out before the actual investigation is done. Researchers use information gathered in pilot studies **to refine or modify the research methodology for a study and to develop large-scale studies**	http://www.mh.state.oh.us/what-we-do/promote/research-and-evaluation/learning-lab/research-glossary.shtml
A small scale study conducted to test the plan and method of a research study.	http://www.umm.edu/nursing/docs/glossary_research_terms.pdf
A small study **carried out before a large-scale study to try out a procedure or to test a principle**	http://www.psych-sci.manchester.ac.uk/actnow/glossary/
An experimental use of a treatment in a small group of patients **to learn if it will be effective and safe on a broad scale**	http://www.lungcanceralliance.org/news/glossary.html
The **initial study examining a new method** or treatment	http://www.cdc.gov/des/consumers/resources/glossary.html#P
A small study often done **to assist the preparation of a larger, more comprehensive investigation**.	http://www.informedesign.umn.edu/Glossary.aspx?id=1952#
Small, preliminary test or trial run of an intervention, or of an evaluation activity such as an instrument or sampling procedure. The results of the pilot are **used to improve the program or evaluation procedure being piloted before it is used on a larger scale**.	http://www.nsf.gov/pubs/2005/nsf0531/nsf0531_6.pdf

*Emphasis is ours

intervention trials [6], cardiovascular trials [7], primary healthcare [8], to mention a few.

Despite their noted importance, the reality is that pilot studies receive little or no attention in scientific research training. Few epidemiology or research textbooks cover the topic with the necessary detail. In fact, we are not aware of any textbook that dedicates a chapter on this issue - many just mention it in passing or provide a cursory coverage of the topic. The objective of this paper is to provide a detailed examination of the key aspects of pilot studies. In the next section, we narrow the focus of our definition of a pilot to phase III trials. Section 3 covers the general reasons for conducting a pilot study. Section 4 deals with the relationships between pilot studies, proof-of-concept studies, and adaptive designs, while section 5 addresses the challenges of pilot studies. Evaluation of a pilot study (i.e. how to determine if a pilot study was successful) is covered in Section 6. We deal with several frequently asked questions about pilot studies in Section 7 using a "question-and-answer" approach. Section 8 covers some ethical aspects related to pilot studies; and in Section 9, we follow the CONSORT format [9] to offer some suggestions on how to report the results of pilot investigations.

2. Narrowing the focus: Pilot studies for randomized studies

Pilot studies can be conducted in both quantitative and qualitative studies. Adopting a similar approach to Lancaster *et al.* [10], we focus on quantitative pilot studies - particularly those done prior to full-scale phase III trials. Phase I trials are non-randomized studies designed to investigate the pharmacokinetics of a drug (i.e. how a drug is distributed and metabolized in the body) including finding a dose that can be tolerated with minimal toxicity. Phase II trials provide preliminary evidence on the clinical efficacy of a drug or intervention. They may or may not be randomized. Phase III trials are randomized studies comparing two or more drugs or intervention strategies to assess efficacy and safety. Phase IV trials, usually done after registration or marketing of a drug, are non-randomized surveillance studies to document experiences (e.g. side-effects, interactions with other drugs, etc) with using the drug in practice.

For the purposes of this paper, our approach to utilizing pilot studies relies on the model for complex interventions advocated by the British Medical Research Council - which explicitly recommends the use of feasibility studies prior to Phase III clinical trials, but stresses the iterative nature of the processes of development, feasibility and piloting, evaluation and implementation [11].

3. Reasons for Conducting Pilot Studies

Van Teijlingen *et al.* [12] and van Teijlingen and Hundley [13] provide a summary of the reasons for performing a pilot study. In general, the rationale for a pilot study can be grouped under several broad classifications - process, resources, management and scientific (see also http://www.childrens-mercy.org/stats/plan/pilot.asp for a different classification):

• *Process:* This assesses the feasibility of the steps that need to take place as part of the main study. Examples include determining recruitment rates, retention rates, etc.

• *Resources:* This deals with assessing time and budget problems that can occur during the main study. The idea is to collect some pilot data on such things as the length of time to mail or fill out all the survey forms.

• *Management:* This covers potential human and data optimization problems such as personnel and data management issues at participating centres.

• *Scientific:* This deals with the assessment of treatment safety, determination of dose levels and response, and estimation of treatment effect and its variance.

Table 2 summarizes this classification with specific examples.

4. Relationships between Pilot Studies, Proof-of-Concept Studies, and Adaptive Designs

A proof-of-concept (PoC) study is defined as a clinical trial carried out to determine if a treatment (drug) is biologically active or inactive [14]. PoC studies usually use surrogate markers as endpoints. In general, they are phase I/II studies - which, as noted above, investigate the safety profile, dose level and response to new drugs [15]. Thus, although designed to inform the planning of phase III trials for registration or licensing of new drugs, PoC studies may not necessarily fit our restricted definition of pilot studies aimed at assessing feasibility of phase III trials as outlined in Section 2.

An adaptive trial design refers to a design that allows modifications to be made to a trial's design or statistical procedures *during* its conduct, with the purpose of efficiently identifying clinical benefits/risks of new drugs or to increase the probability of success of clinical development [16]. The adaptations can be prospective (e.g. stopping a trial early due to safety or futility or efficacy at interim analysis); concurrent (e.g. changes in eligibility criteria, hypotheses or study endpoints) or retrospective (e.g. changes to statistical analysis plan prior to locking database or revealing treatment codes to trial investigators or patients). Piloting is normally built into adaptive trial designs by determining a *priori* decision rules to guide the adaptations based on cumulative data. For example, data from interim analyses could be used to refine sample size calculations [17,18]. This approach is routinely used in internal pilot studies - which are primarily designed to inform sample size calculation for the main study, with recalculation of the sample size as the key adaptation. Unlike other phase III pilots, an internal pilot investigation does not usually address any other feasibility aspects - because it is essentially part of the main study [10,19,20]..

Nonetheless, we need to emphasize that whether or not a study is a pilot, depends on its objectives. An adaptive method is used as a strategy to reach that objective. Both a pilot and a non-pilot could be adaptive.

5. Challenges of and Common Misconceptions about Pilot Studies

Pilot studies can be very informative, not only to the researchers conducting them but also to others doing similar work. However, many of them never get published, often because of the way the results are presented [13]. Quite often the emphasis is wrongly placed on statistical significance, not on feasibility - which is the main focus of the pilot study. Our experience in reviewing submissions to a research ethics board also shows that most of the pilot projects are not well designed: i.e. there are no clear feasibility objectives; no clear analytic plans; and certainly no clear criteria for success of feasibility.

In many cases, pilot studies are conducted to generate data for sample size calculations. This seems especially sensible in situations where there are no data from previous studies to inform this process. However, it can be dangerous to use pilot studies to estimate treatment effects, as such estimates may be unrealistic/biased because of the limited sample sizes. Therefore if not used cautiously, results of pilot studies can potentially mislead sample size or power calculations [21] – particularly if the pilot study was done to see if there is likely to be a treatment effect in the main study. In section 6, we provide guidance on how to proceed with caution in this regard.

There are also several misconceptions about pilot studies. Below are some of the common reasons that researchers have put forth for calling their study a pilot.

The first common reason is that a pilot study is a small single-centre study. For example, researchers often state lack of resources for a large multi-centre study as a reason for doing a pilot. The second common reason is that a pilot investigation is a small study that is similar in size to someone else's published study. In reviewing submissions to a research ethics board, we have come across sentiments such as

• *So-and-so did a similar study with 6 patients and got statistical significance - ours uses 12 patients (double the size)!*

• *We did a similar pilot before (and it was published!)*

The third most common reason is that a pilot is a small study done by a student or an intern - which can be completed quickly and does not require funding. Specific arguments include

• *I have funding for 10 patients only;*

• *I have limited seed (start-up) funding;*

• *This is just a student project!*

Table 2 Reasons for conducting pilot studies

Main Reason	Examples
Process: This assesses the feasibility of the processes that are key to the success of the main study	• Recruitment rates
	• Retention rates
	• Refusal rates
	• Failure/success rates
	• (Non)compliance or adherence rates
	• eligibility criteria
	- Is it obvious who meets and who does not meet the eligibility requirements?
	- Are the eligibility criteria sufficient or too restrictive?
	• Understanding of study questionnaires or data collection tools:
	- Do subjects provide no answer, multiple answers, qualified answers, or unanticipated answers to study questions?
Resources: This deals with assessing time and resource problems that can occur during the main study	• Length of time to fill out all the study forms
	• Determining capacity:
	- Will the study participants overload your phone lines or overflow your waiting room?
	• Determining process time
	- How much time does it take to mail out a thousand surveys?
	• Is the equipment readily available when and where it is needed?
	• What happens when it breaks down or gets stolen?
	• Can the software used for capturing data read and understand the data?
	• Determining centre willingness and capacity
	- Do the centres do what they committed to doing?
	- Do investigators have the time to Perform the tasks they committed to doing?
	- Are there any capacity issues at each participating centre?
Management: This covers potential human and data management problems	• What are the challenges that participating centres have with managing the study?
	• What challenges do study personnel have?
	• Is there enough room on the data collection form for all of the data you receive?
	• Are there any problems entering data into the computer?
	• Can data coming from different sources be matched?
	• Were any important data values forgotten about?
	• Do data show too much or too little variability?
Scientific: This deals with the assessment of treatment safety, dose, response, effect and variance of the effect	• Is it safe to use the study drug/intervention?
	• What is the safe dose level?
	• Do patients respond to the drug?
	• What is the estimate of the treatment effect?
	• What is the estimate of the variance of the treatment effect?

• *My supervisor (boss) told me to do it as a pilot.*

None of the above arguments qualifies as sound reasons for calling a study a pilot. A study should only be conducted if the results will be informative; studies conducted for the reasons above may result in findings of limited utility, which would be a waste of the researchers' and participants' efforts. The focus of a pilot study should be on assessment of feasibility, unless it was powered appropriately to assess statistical significance. Further, there is a vast number of poorly designed and reported studies. Assessment of the quality of a published report may be helpful to guide decisions of whether the report should be used to guide planning or designing of new studies. Finally, if a trainee or researcher is assigned a project as a pilot it is important to discuss how the results will inform the planning of the main study. In addition, clearly defined feasibility

objectives and rationale to justify piloting should be provided.

Sample Size for Pilot Studies

In general, sample size calculations may not be required for some pilot studies. It is important that the sample for a pilot be representative of the target study population. It should also be based on the same inclusion/exclusion criteria as the main study. As a rule of thumb, a pilot study should be large enough to provide useful information about the aspects that are being assessed for feasibility. Note that *PoC studies* require sample size estimation based on surrogate markers [22], but they are usually not powered to detect meaningful differences in clinically important endpoints. The sample used in the pilot may be included in the main study, but caution is needed to ensure the key features of the main study are preserved in the pilot (e.g. blinding in randomized controlled trials). We recommend if any pooling of pilot and main study data is considered, this should be planned beforehand, described clearly in the protocol with clear discussion of the statistical consequences and methods. The goal is to avoid or minimize the potential bias that may occur due to multiple testing issues or any other opportunistic actions by investigators. In general, pooling when done appropriately can increase the efficiency of the main study [23].

As noted earlier, a carefully designed pilot study may be used to generate information for sample size calculations. Two approaches may be helpful to optimize information from a pilot study in this context: First, consider eliciting qualitative data to supplement the quantitative information obtained in the pilot. For example, consider having some discussions with clinicians using the approach suggested by Lenth [24] to illicit additional information on possible effect size and variance estimates. Second, consider creating a sample size table for various values of the effect or variance estimates to acknowledge the uncertainty surrounding the pilot estimates.

In some cases, one could use a **confidence interval [CI] approach** to estimate the sample size required to establish feasibility. For example, suppose we had a pilot trial designed primarily to determine adherence rates to the standardized risk assessment form to enhance venous thromboprophylaxis in hospitalized patients. Suppose it was also decided *a priori* that the criterion for success would be: the main trial would be '*feasible*' if the risk assessment form is completed for ≥ 70% of eligible hospitalized patients.

Using a 95% CI for the proportion of eligible patients who complete the assessment form, a margin of error (ME) of 0.05, a lower bound of this CI of 0.70, and an expected completion rate of 75% based on an educated guess, the required sample for the pilot study would be at least 75 patients. This calculation is based on a common formula for obtaining a 95% CI for a single proportion: $p \pm 1.96 \sqrt{\frac{p(1-p)}{n}}$ where "p" is the prior estimate of the proportion of interest and "n" is the sample size.

6. How to Interpret the Results of a Pilot Study: Criteria for Success

It is always important to state the criteria for success of a pilot study. The criteria should be based on the primary feasibility objectives. These provide the basis for interpreting the results of the pilot study and determining whether it is feasible to proceed to the main study. In general, the outcome of a pilot study can be one of the following: (i) *Stop* - main study not feasible; (ii) *Continue, but modify protocol* - feasible with modifications; (iii) *Continue without modifications, but monitor closely* - feasible with close monitoring and (iv) *Continue without modifications* - feasible as is.

For example, the Prophylaxis of Thromboembolism in Critical Care Trial (PROTECT) was designed to assess the feasibility of a large-scale trial with the following criteria for determining success [25]:

- *98.5% of patients had to receive study drug within 12 hours of randomization;*
- *91.7% of patients had to receive every scheduled dose of the study drug in a blinded manner;*
- *90% or more of patients had to have lower limb compression ultrasounds performed at the specified times; and*
- *> 90% of necessary dose adjustments had to have been made appropriately in response to pre-defined laboratory criteria.*

In a second example, the PeriOperative Epidural Trial (POET) Pilot Study was designed to assess the feasibility of a large, multicentre trial with the following criteria for determining success [26]:

- *one subject per centre per week (i.e., 200 subjects from four centres over 50 weeks) can be recruited;*
- *at least 70% of all eligible patients can be recruited;*
- *no more than 5% of all recruited subjects crossed over from one modality to the other; and*
- *complete follow-up in at least 95% of all recruited subjects.*

7. Frequently asked questions about pilot studies

In this Section, we offer our thoughts on some of the frequently asked questions about pilot studies. These

could be helpful to not only clinicians and trainees, but to anyone who is interested in health research.

• ***Can I publish the results of a pilot study?***

- Yes, every attempt should be made to publish.

• ***Why is it important to publish the results of pilot studies?***

- To provide information about feasibility to the research community to save resources being unnecessarily spent on studies that may not be feasible. Further, having such information can help researchers to avoid duplication of efforts in assessing feasibility.

- Finally, researchers have an ethical and scientific obligation to attempt publishing the results of every research endeavor. However, our focus should be on feasibility goals. Emphasis should not be placed on statistical significance when pilot studies are not powered to detect minimal clinically important differences. Such studies typically do not show statistically significant results - remember that underpowered studies (with no statistically significant results) are inconclusive, not negative since "no evidence of effect" is not "evidence of no effect" [27].

• ***Can I combine data from a pilot with data from the main study?***

- Yes, provided the sampling frame and methodologies are the same. This can increase the efficiency of the main study - see Section 5.

• ***Can I combine the results of a pilot with the results of another study or in a meta-analysis?***

- Yes, provided the sampling frame and methodologies are the same.

- No, if the main study is reported and it includes the pilot study.

• ***Can the results of the pilot study be valid on their own, without existence of the main study***

- Yes, if the results show that it is not feasible to proceed to the main study or there is insufficient funding.

• ***Can I apply for funding for a pilot study?***

- Yes. Like any grant, it is important to justify the need for piloting.

- The pilot has to be placed in the context of the main study.

• ***Can I randomize patients in a pilot study?***

- Yes. For a phase III pilot study, one of the goals could be to assess how a randomization procedure might work in the main study or whether the idea of randomization might be acceptable to patients [10]. In general, it is always best for a pilot to maintain the same design as the main study.

• ***How can I use the information from a pilot to estimate the sample size?***

- Use with caution, as results from pilot studies can potentially mislead sample size calculations.

- Consider supplementing the information with qualitative discussions with clinicians - see section 5; and

- Create a sample size table to acknowledge the uncertainty of the pilot information - see section 5.

• ***Can I use the results of a pilot study to treat my patients?***

- Not a good idea!

- Pilot studies are primarily for assessing feasibility.

• ***What can I do with a failed or bad pilot study?***

- No study is a complete failure; it can always be used as bad example! However, it is worth making clear that a pilot study that shows the main study is not likely to be feasible is not a failed (pilot) study. In fact, it is a success - because you avoided wasting scarce resources on a study destined for failure!

8. Ethical Aspects of Pilot Studies

Halpern *et al.* [28] stated that conducting underpowered trials is unethical. However, they proposed that underpowered trials are ethical in two situations: (i) small trials of interventions for rare diseases – which require documenting explicit plans for including results with those of similar trials in a prospective meta-analysis; (ii) early-phase trials in the development of drugs or devices - provided they are adequately powered for defined purposes other than randomized treatment comparisons. Pilot studies of phase III trials (dealing with common diseases) are not addressed in their proposal. It is therefore prudent to ask: *Is it ethical to conduct a study whose feasibility can not be guaranteed (i.e. with a high probability of success)?*

It seems unethical to consider running a phase III study without having sufficient data or information about the feasibility. In fact, most granting agencies often require data on feasibility as part of their assessment of the scientific validity for funding decisions.

There is however one important ethical aspect about pilot studies that has received little or no attention from researchers, research ethics boards and ethicists alike. This pertains to the issue of the obligation that researchers have to patients or participants in a trial to disclose the feasibility nature of pilot studies. This is essential given that some pilot studies may not lead to further studies. A review of the commonly cited research ethics guidelines - the Nuremburg Code [29], Helsinki Declaration [30], the Belmont Report [31], ICH Good Clinical Practice [32], and the International Ethical Guidelines for Biomedical Research Involving Human Subjects [33] - shows that pilot studies are not addressed in any of these guidelines. Canadian researchers are also encouraged to follow the Tri-Council Policy Statement (TCPS) [34] - it too does not address how pilot studies need to be approached. It seems to us that given the special nature of feasibility or pilot studies, the

disclosure of their purpose to study participants requires special wording - that informs them of the definition of a pilot study, the feasibility objectives of the study, and also clearly defines the criteria for success of feasibility. To fully inform participants, we suggest using the following wording in the consent form:

> "*The overall purpose of this pilot study is to assess the feasibility of conducting a large study to [state primary objective of the main study]. A feasibility or pilot study is a study that... [state a general definition of a feasibility study]. The specific feasibility objectives of this study are ... [state the specific feasibility objectives of the pilot study]. We will determine that it is feasible to carry on the main study if ... [state the criteria for success of feasibility].*"

9. Recommendation for Reporting the Results of Pilot Studies

Adopted from the CONSORT Statement [9], Table 3 provides a checklist of items to consider including in a report of a pilot study.

Title and abstract

Item #1: The title or abstract should indicate that the study is a "pilot" or "feasibility"

As a number one summary of the contents of any report, it is important for the title to clearly indicate that the report is for a pilot or feasibility study. This would also be helpful to other researchers during electronic information search about feasibility issues. Our quick search of PUBMED [on July 13, 2009], using the terms "pilot" OR "feasibility" OR "proof-of-concept" for revealed 24423 (16%) hits of studies that had these terms in the title or abstract compared with 149365 hits that had these terms anywhere in the text.

Background

Item #2: Scientific background for the main study and explanation of rationale for assessing feasibility through piloting

The rationale for initiating a pilot should be based on the need to assess feasibility for the main study. Thus, the background of the main study should clearly describe what is known or not known about important feasibility aspects to provide context for piloting.

Methods

Item #3: Participants and setting of the study

The description of the inclusion-exclusion or eligibility criteria for participants should be the same as in the main study. The settings and locations where the data were collected should also be clearly described.

Item #4: Interventions

Precise details of the interventions intended for each group and how and when they were actually administered (if applicable) - state clearly if any aspects of the intervention are assessed for feasibility.

Item #5: Objectives

State the specific scientific primary and secondary objectives and hypotheses for the main study and the specific feasibility objectives. It is important to clearly indicate the feasibility objectives as the primary focus for the pilot.

Item #6: Outcomes

Clearly define primary and secondary outcome measures for the main study. Then, clearly define the feasibility outcomes and how they were operationalized - these should include key elements such as recruitment rates, consent rates, completion rates, variance estimates, etc. In some cases, a pilot study may be conducted with the aim to determine a suitable (clinical or surrogate) endpoint for the main study. In such a case, one may not be able to define the primary outcome of the main study until the pilot is finished. However, it is important that determining the primary outcome of the main study be clearly stated as part of feasibility outcomes.

Item #7: Sample Size

Describe how sample size was determined. If the pilot is a proof-of-concept study, is the sample size calculated based on primary/key surrogate marker(s)? In general if the pilot is for a phase III study, there may be no need for a formal sample size calculation. However, the confidence interval approach may be used to calculate and justify the sample size based on key feasibility objective (s).

Item #8: Feasibility criteria

Clearly describe the criteria for assessing success of feasibility - these should be based on the feasibility objectives.

Item #9: Statistical Analysis

Describe the statistical methods for the analysis of primary and secondary feasibility outcomes.

Item #10: Ethical Aspects

State whether the study received research ethics approval. Describe how informed consent was handled - given the feasibility nature of the study.

Results

Item #11: Participant Flow

Describe the flow of participants through each stage of the study (use of a flow-diagram is strongly recommended – see CONSORT [9] for a template). Describe protocol deviations from pilot study as planned with reasons for deviations. State the number of exclusions at each stage and corresponding reasons for exclusions.

Item #12: Recruitment

Report the dates defining the periods of recruitment and follow-up.

Item #13: Baseline Data

Report the baseline demographic and clinical characteristics of the participants.

Table 3 Pilot Study - Checklist: Items to include when reporting a pilot study

PAPER SECTION	Item	Descriptor	Reported on Page #
TITLE *and* ABSTRACT	1	Does the title or abstract indicate that the study is a "pilot"?	
INTRODUCTION			
Background	2	Scientific background for the main study and explanation of rationale for assessing feasibility through piloting	
METHODS			
Participants and setting	3	• Eligibility criteria for participants in the pilot study (these should be the same as in the main study – if different, state the differences)	
		• The settings and locations where the data were collected	
Interventions	4	Provide precise details of the interventions intended for each group and how and when they were actually administered (if applicable) – state clearly if any aspects of the intervention are assessed for feasibility	
Objectives	5	• Specific scientific objectives and hypotheses for the main study	
		• Specific feasibility objectives	
Outcomes	6	• Clearly defined primary and secondary outcome measures for the main study	
		• Clearly define the feasibility outcomes and how they were operationalized – these should include key elements such as recruitment rates, consent rates, completion rates, variance estimates, etc	
Sample size	7	Describe how sample size was determined	
		• In general for a pilot of a phase III trial, there is no need for a formal sample size calculation. However, confidence interval approach may be used to calculate and justify the sample size based on key feasibility objective(s).	
Feasibility Criteria	8	Clearly describe the criteria for assessing success of feasibility – these should be based on the feasibility objectives	
Statistical Methods	9	Describe the statistical methods for the analysis of primary and secondary feasibility outcomes	
Ethical Aspects	10	• State whether the study received research ethics approval	
		• State how informed consent was handled – given the feasibility nature of the study	
RESULTS			
Participant flow	11	Flow of participants through each stage (a flow-chart is strongly recommended).	
		• Describe protocol deviations from pilot study as planned, together with reasons	
		• State the number of exclusions at each stage and reasons for exclusions	
Recruitment	12	Report the dates defining the periods of recruitment and follow-up	
Baseline data	13	Report the baseline demographic and clinical characteristics of the participants	
Outcomes and estimation	14	For each primary and secondary feasibility outcome, report the point estimate of effect and its precision (*e.g.*, 95% confidence interval [CI]) – if applicable	
DISCUSSION			
Interpretation	15	Interpretation of the results should focus on feasibility, taking into account	
		• the stated criteria for success of feasibility;	
		• study hypotheses, sources of potential bias or imprecision – given the feasibility nature of the study	
		• the dangers associated with multiplicity of analyses and outcomes	
Generalizability	16	Generalizability (external validity) of the feasibility. State clearly what modifications in the design of the main study (if any) would be necessary to make it feasible	
Overall evidence of feasibility	17	General interpretation of the results in the context of current evidence of feasibility	
		• Focus should be on feasibility	

Item #14: Outcomes and Estimation

For each primary and secondary feasibility outcomes, report the point estimate of effect and its precision (*e.g.*, 95% CI) - if applicable.

Discussion

Item # 15: Interpretation

Interpretation of the results should focus on feasibility, taking into account the stated criteria for success of feasibility, study hypotheses, sources of potential bias or imprecision (given the feasibility nature of the study) and the dangers associated with multiplicity - repeated testing on multiple outcomes.

Item #16: Generalizability

Discuss the generalizability (external validity) of the feasibility aspects observed in the study. State clearly what modifications in the design of the main study (if any) would be necessary to make it feasible.

Item #17: Overall evidence of feasibility

Discuss the general results in the context of overall evidence of feasibility. It is important that the focus be on feasibility.

9. Conclusions

Pilot or vanguard studies provide a good opportunity to assess feasibility of large full-scale studies. Pilot studies are the best way to assess feasibility of a large expensive full-scale study, and in fact are an almost essential prerequisite. Conducting a pilot prior to the main study can enhance the likelihood of success of the main study and potentially help to avoid doomed main studies. Pilot studies should be well designed with clear feasibility objectives, clear analytic plans, and explicit criteria for determining success of feasibility. They should be used cautiously for determining treatment effects and variance estimates for power or sample size calculations. Finally, they should be scrutinized the same way as full scale studies, and every attempt should be taken to publish the results in peer-reviewed journals.

Acknowledgements

Dr Lehana Thabane is clinical trials mentor for the Canadian Institutes of Health Research. We thank the reviewers for insightful comments and suggestions which led to improvements in the manuscript.

Author details

[1]Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton ON, Canada. [2]Biostatistics Unit, St Joseph's Healthcare Hamilton, Hamilton ON, Canada. [3]Department of Medical Affairs, GlaxoSmithKline Inc., Mississauga ON, Canada. [4]Department of Medicine, Division of Gastroenterology, McMaster University, Hamilton ON, Canada. [5]Department of Kinesiology, University of Waterloo, Waterloo ON, Canada.

Authors' contributions

LT drafted the manuscript. All authors reviewed several versions of the manuscript, read and approved the final version.

Competing interests

The authors declare that they have no competing interests.

Received: 9 August 2009
Accepted: 6 January 2010 Published: 6 January 2010

References

1. Waite M: *Concise Oxford Thesaurus* Oxford, England: Oxford University Press, 2 2002.
2. Last JM, editor: *A Dictionary of Epidemiology* Oxford University Press, 4 2001.
3. Everitt B: *Medical Statistics from A to Z: A Guide for Clinicians and Medical Students* Cambridge University Press: Cambridge, 2 2006.
4. Tavel JA, Fosdick L, ESPRIT Vanguard Group. ESPRIT Executive Committee: **Closeout of four phase II Vanguard trials and patient rollover into a large international phase III HIV clinical endpoint trial.** *Control Clin Trials* 2001, **22**:42-48.
5. Arnold DM, Burns KE, Adhikari NK, Kho ME, Meade MO, Cook DJ: **The design and interpretation of pilot trials in clinical research in critical care.** *Crit Care Med* 2009, **37(Suppl 1)**:69-74.
6. **Computerization of Medical Practice for the Enhancement of Therapeutic Effectiveness.** http://www.compete-study.com/index.htm , Last accessed August 8, 2009.
7. **Heart Outcomes Prevention Evaluation Study.** http://www.ccc.mcmaster.ca/hope.htm, Last accessed August 8, 2009.
8. **Cardiovascular Health Awareness Program.** http://www.chapprogram.ca/resources.html, Last accessed August 8, 2009.
9. Moher D, Schulz KF, Altman DG, CONSORT Group (Consolidated Standards of Reporting Trials): **The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials.** *J Am Podiatr Med Assoc* 2001, **91**:437-442.
10. Lancaster GA, Dodd S, Williamson PR: **Design and analysis of pilot studies: recommendations for good practice.** *J Eval Clin Pract* 2004, **10**:307-12.
11. Craig N, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M: **Developing and evaluating complex interventions: the new Medical Research Council guidance.** *BMJ* 2008, **337**:a1655.
12. Van Teijlingen ER, Rennie AM, Hundley V, Graham W: **The importance of conducting and reporting pilot studies: the example of the Scottish Births Survey.** *J Adv Nurs* 2001, **34**:289-295.
13. Van Teijlingen ER, Hundley V: **The Importance of Pilot Studies.** *Social Research Update* 2001, 35http://sru.soc.surrey.ac.uk/SRU35.html.
14. Lawrence Gould A: **Timing of futility analyses for 'proof of concept' trials.** *Stat Med* 2005, **24**:1815-1835.
15. Fardon T, Haggart K, Lee DK, Lipworth BJ: **A proof of concept study to evaluate stepping down the dose of fluticasone in combination with salmeterol and tiotropium in severe persistent asthma.** *Respir Med* 2007, **101**:1218-1228.
16. Chow SC, Chang M: **Adaptive design methods in clinical trials - a review.** *Orphanet J Rare Dis* 2008, **3**:11.
17. Gould AL: **Planning and revising the sample size for a trial.** *Stat Med* 1995, **14**:1039-1051.
18. Coffey CS, Muller KE: **Properties of internal pilots with the univariate approach to repeated measures.** *Stat Med* 2003, **22**:2469-2485.
19. Zucker DM, Wittes JT, Schabenberger O, Brittain E: **Internal pilot studies II: comparison of various procedures.** *Statistics in Medicine* 1999, **18**:3493-3509.
20. Kieser M, Friede T: **Re-calculating the sample size in internal pilot designs with control of the type I error rate.** *Statistics in Medicine* 2000, **19**:901-911.
21. Kraemer HC, Mintz J, Noda A, Tinklenberg J, Yesavage JA: **Caution regarding the use of pilot studies to guide power calculations for study proposals.** *Arch Gen Psychiatry* 2006, **63**:484-489.
22. Yin Y: **Sample size calculation for a proof of concept study.** *J Biopharm Stat* 2002, **12**:267-276.
23. Wittes J, Brittain E: **The role of internal pilot studies in increasing the efficiency of clinical trials.** *Stat Med* 1990, **9**:65-71.
24. Lenth R: **Some Practical Guidelines for Effective Sample Size Determination.** *The American Statistician* 2001, **55**:187-193.
25. Cook DJ, Rocker G, Meade M, Guyatt G, Geerts W, Anderson D, Skrobik Y, Hebert P, Albert M, Cooper J, Bates S, Caco C, Finfer S, Fowler R, Freitag A, Granton J, Jones G, Langevin S, Mehta S, Pagliarello G, Poirier G, Rabbat C, Schiff D, Griffith L, Crowther M, PROTECT Investigators. Canadian Critical Care Trials Group: **Prophylaxis of Thromboembolism in Critical Care (PROTECT) Trial: a pilot study.** *J Crit Care* 2005, **20**:364-372.
26. Choi PT, Beattie WS, Bryson GL, Paul JE, Yang H: **Effects of neuraxial blockade may be difficult to study using large randomized controlled**

trials: the PeriOperative Epidural Trial (POET) Pilot Study. *PLoS One* 2009, **4(2)**:e4644.
27. Altman DG, Bland JM: **Absence of evidence is not evidence of absence.** *BMJ* 1995, **311**:485.
28. Halpern SD, Karlawish JH, Berlin JA: **The continuing unethical conduct of underpowered clinical trials.** *JAMA* 2002, **288**:358-362.
29. **The Nuremberg Code, Research ethics guideline 2005.** http://www.hhs.gov/ohrp/references/nurcode.htm , Last accessed August 8, 2009.
30. **The Declaration of Helsinki, Research ethics guideline.** http://www.wma.net/en/30publications/10policies/b3/index.html, Last accessed December 22, 2009.
31. **The Belmont Report, Research ethics guideline.** http://ohsr.od.nih.gov/guidelines/belmont.html, Last accessed August 8, 2009.
32. **The ICH Harmonized Tripartite Guideline-Guideline for Good Clinical Practice.** http://www.gcppl.org.pl/ma_struktura/docs/ich_gcp.pdf, Last accessed August 8, 2009.
33. **The International Ethical Guidelines for Biomedical Research Involving Human Subjects.** http://www.fhi.org/training/fr/Retc/pdf_files/cioms.pdf, Last accessed August 8, 2009.
34. **Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans, Government of Canada.** http://www.pre.ethics.gc.ca/english/policystatement/policystatement.cfm, Last accessed August 8, 2009.

Pre-publication history

The pre-publication history for this paper can be accessed here:http://www.biomedcentral.com/1471-2288/10/1/prepub

doi:10.1186/1471-2288-10-1
Cite this article as: Thabane *et al.*: A tutorial on pilot studies: the what, why and how. *BMC Medical Research Methodology* 2010 **10**:1.

PLOS ONE

RESEARCH ARTICLE

Defining Feasibility and Pilot Studies in Preparation for Randomised Controlled Trials: Development of a Conceptual Framework

Sandra M. Eldridge[1]*, Gillian A. Lancaster[2], Michael J. Campbell[3], Lehana Thabane[4], Sally Hopewell[5], Claire L. Coleman[1], Christine M. Bond[6]

1 Centre for Primary Care and Public Health, Queen Mary University of London, London, United Kingdom, **2** Department of Mathematics and Statistics, Lancaster University, Lancaster, Lancashire, United Kingdom, **3** School of Health and Related Research, University of Sheffield, Sheffield, South Yorkshire, United Kingdom, **4** Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada, **5** Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, Oxfordshire, United Kingdom, **6** Centre of Academic Primary Care, University of Aberdeen, Aberdeen, Scotland, United Kingdom

* s.eldridge@qmul.ac.uk

OPEN ACCESS

Citation: Eldridge SM, Lancaster GA, Campbell MJ, Thabane L, Hopewell S, Coleman CL, et al. (2016) Defining Feasibility and Pilot Studies in Preparation for Randomised Controlled Trials: Development of a Conceptual Framework. PLoS ONE 11(3): e0150205. doi:10.1371/journal.pone.0150205

Editor: Chiara Lazzeri, Azienda Ospedaliero-Universitaria Careggi, ITALY

Received: August 13, 2015

Accepted: February 10, 2016

Published: March 15, 2016

Data Availability Statement: Due to a requirement by the ethics committee that the authors specified when the data will be destroyed, the authors are not able to give unlimited access to the Delphi study quantitative data. These data are available from Professor Sandra Eldridge. Data will be available upon request to all interested researchers. Qualitative data from the Delphi study are not available because the authors do not have consent from participants for wider distribution of this more sensitive data.

Funding: The authors received small grants from Queen Mary University of London (£7495), University

Abstract

We describe a framework for defining pilot and feasibility studies focusing on studies conducted in preparation for a randomised controlled trial. To develop the framework, we undertook a Delphi survey; ran an open meeting at a trial methodology conference; conducted a review of definitions outside the health research context; consulted experts at an international consensus meeting; and reviewed 27 empirical pilot or feasibility studies. We initially adopted mutually exclusive definitions of pilot and feasibility studies. However, some Delphi survey respondents and the majority of open meeting attendees disagreed with the idea of mutually exclusive definitions. Their viewpoint was supported by definitions outside the health research context, the use of the terms 'pilot' and 'feasibility' in the literature, and participants at the international consensus meeting. In our framework, pilot studies are a subset of feasibility studies, rather than the two being mutually exclusive. A feasibility study asks whether something can be done, should we proceed with it, and if so, how. A pilot study asks the same questions but also has a specific design feature: in a pilot study a future study, or part of a future study, is conducted on a smaller scale. We suggest that to facilitate their identification, these studies should be clearly identified using the terms 'feasibility' or 'pilot' as appropriate. This should include feasibility studies that are largely qualitative; we found these difficult to identify in electronic searches because researchers rarely used the term 'feasibility' in the title or abstract of such studies. Investigators should also report appropriate objectives and methods related to feasibility; and give clear confirmation that their study is in preparation for a future randomised controlled trial designed to assess the effect of an intervention.

of Sheffield (£8000), NIHR RDS London (£2000), NIHR RDS South East (£2400), Chief Scientist Office Scotland (£1000). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: All authors have completed the ICMJE disclosure form at http://www.icmje.org/coi_disclosure.pdf and declare support from the following organisations that might have an interest in the submitted work – Queen Mary University of London, Sheffield University, NIHR, Chief Scientist Office Scotland; financial relationships with NIHR, MRC, EC FP7, Canadian Institute for Health Research, Wiley, who might have an interest in the submitted work in the previous three years. No other relationships or activities have influenced the submitted work. This does not alter the authors' adherence to PLOS ONE policies on sharing data and materials.

Introduction

There is a large and growing number of studies in the literature that authors describe as feasibility or pilot studies. In this paper we focus on feasibility and pilot studies conducted in preparation for a future definitive randomised controlled trial (RCT) that aims to assess the effect of an intervention. We are primarily concerned with stand-alone studies that are completed before the start of such a definitive RCT, and do not specifically cover internal pilot studies which are designed as the early stage of a definitive RCT; work on the conduct of internal pilot studies is currently being carried out by the UK MRC Network of Hubs for Trials Methodology Research. One motivating factor for the work reported in this paper was the inconsistent use of terms. For example, in the context of RCTs 'pilot study' is sometimes used to refer to a study addressing feasibility in preparation for a larger RCT, but at other times it is used to refer to a small scale, often opportunistic, RCT which assesses efficacy or effectiveness.

A second, related, motivating factor was the lack of agreement in the research community about the use of the terms 'pilot' and 'feasibility' in relation to studies conducted in preparation for a future definitive RCT. In a seminal paper in 2004 reviewing the literature in relation to pilot and feasibility studies conducted in preparation for an RCT [1], Lancaster *et al* reported that they could find no formal guidance as to what constituted a pilot study. In the updated UK Medical Research Council (MRC) guidance on designing and evaluating complex interventions published four years later, feasibility and pilot studies are explicitly recommended, particularly in relation to identifying problems that might occur in an ensuing RCT of a complex intervention [2]. However, while the guidance suggests possible aims of such studies, for example, testing procedures for their acceptability, estimating the likely rates of recruitment and retention of subjects, and the calculation of appropriate sample sizes, no explicit definitions of a 'pilot study' or 'feasibility study' are provided. In 2010, Thabane and colleagues presented a number of definitions of pilot studies taken from various health related websites [3]. While these definitions vary, most have in common the idea of conducting a study in advance of a larger, more comprehensive, investigation. Thabane *et al* also considered the relationship between pilot and feasibility, suggesting that feasibility should be the main emphasis of a pilot study and that 'a pilot study is synonymous with a feasibility study intended to guide the planning of a large scale investigation'. However, at about the same time, the UK National Institute for Health Research (NIHR) developed definitions of pilot and feasibility studies that are mutually exclusive, suggesting that feasibility studies occurred slightly earlier in the research process and that pilot studies are 'a version of the main study that is run in miniature to test whether the components of the main study can all work together'. Arain *et al*. felt that the NIHR definitions were helpful, and showed that studies identified using the keyword 'feasibility' had different characteristics from those identified as 'pilot' studies [4]. The NIHR wording for pilot studies has been changed more recently to 'a smaller version of the main study used to test whether the components of the main study can all work together' (Fig 1). Nevertheless, it still contrasts with the MRC framework guidance that explicitly states: 'A pilot study need not be a "scale model" of the planned main-stage evaluation, but should address the main uncertainties that have been identified in the development work' [2]. These various, sometimes conflicting, approaches to the interpretation of the terms 'pilot' and 'feasibility' exemplify differences in current usage and opinion in the research community.

While lack of agreement about definitions may not necessarily affect research quality, it can become problematic when trying to develop guidance for research conduct because of the need for clarity over what the guidance applies to and therefore what it should contain. Previous research has identified weaknesses in the reporting and conduct of pilot and feasibility studies [1, 3, 4, 7], particularly in relation to studies conducted in preparation for a future definitive

Feasibility Studies are pieces of research done before a main study in order to answer the question "Can this study be done?". They are used to estimate important parameters that are needed to design the main study. For instance:

- standard deviation of the outcome measure, which is needed in some cases to estimate sample size;
- willingness of participants to be randomised;
- willingness of clinicians to recruit participants;
- number of eligible patients, carers or other appropriate participants
- characteristics of the proposed outcome measure and in some cases feasibility studies might involve designing a suitable outcome measure;
- follow-up rates, response rates to questionnaires, adherence/compliance rates, ICCs in cluster trials, etc.
- availability of data needed or the usefulness and limitations of a particular database
- time needed to collect and analyse data

Pilot studies are a smaller version of the main study used to test whether the components of the main study can all work together. It is focused on the processes of the main study, for example to ensure that recruitment, randomisation, treatment, and follow-up assessments all run smoothly. It resembles the main study in many respects, including an assessment of the primary outcome. In some cases, this will be the first phase of the substantive study and data from the pilot phase may contribute to the final analysis; this can be referred to as an internal pilot. Or, at the end of the pilot study, the data may be analysed and set aside, a so-called external pilot.

Fig 1. NIHR definitions [5, 6].

doi:10.1371/journal.pone.0150205.g001

RCT assessing the effect of an intervention or therapy. While undertaking research to develop guidance to address some of the weaknesses in reporting these studies, we became convinced by the current interest in this area, the lack of clarity, and the differences of opinion in the research community, that a re-evaluation of the definitions of pilot and feasibility studies was needed. This paper describes the process and results of this re-evaluation and suggests a conceptual framework within which researchers can operate when designing and reporting pilot/feasibility studies. Since our work on reporting guidelines focused specifically on pilot and feasibility studies in preparation for an RCT assessing the effect of some intervention or therapy, we restrict our re-evaluation to these types of pilot and feasibility studies.

Methods

The process of developing and validating the conceptual framework for defining pilot and feasibility studies was, to a large extent, integral to the development of our reporting guidelines, the core components of which were a large Delphi study and an international expert consensus meeting focused on developing an extension of the 2010 CONSORT statement for RCTs [8] to randomised pilot studies. The reporting guidelines, Delphi study and consensus meeting are therefore referred to in this paper. However, the reporting guidelines will be reported separately; this paper focuses on our conceptual framework.

Developing a conceptual framework—Delphi study

Following research team discussion of our previous experience with, and research on, pilot and feasibility studies we initially produced mutually exclusive definitions of pilot and feasibility studies based on, but not identical to, the definitions used by the NIHR. We drew up two draft reporting checklists based on the 2010 CONSORT statement [8], one for what we had defined as feasibility studies and one for what we had defined as pilot studies. We constructed a Delphi survey, administered on-line by Clinvivo [9], to obtain consensus on checklist items for inclusion in a reporting guideline, and views on the definitions. Following user-testing of a draft

.....Thus we are focusing on individual studies conducted in preparation for randomised controlled trials.

Many such studies involve implementing a reasonably well-defined intervention on a relatively small scale. These may or may not involve randomisation. Such studies, whether randomised or not, are relatively common in the literature..... We refer to these as pilot studies for a randomised controlled trial and formally define them as: *A pilot study for a randomised controlled trial focuses on the integrity of the study protocol for the main trial. It asks the question 'Will this protocol work, if not why not and how should it be changed?' It may therefore resemble the main trial in many respects, particularly in study design and conduct. However the aim is not to test the effectiveness of the intervention but to assess the feasibility of the protocol. Within this general aim investigators may focus on specific objectives which address key areas of uncertainty.*

Another type of study involves assessing the feasibility of the some aspect of a trial without implementing a well-defined intervention..... We refer to these studies as feasibility studies for randomised controlled trials but do not consider them to be pilot studies. We formally define these studies as "*A feasibility study for a randomised controlled trial answers the question "Can these potential parts of a trial be done?" It has a clearly defined objective or set of objectives. These objectives need to be met for the next stage of study development to go ahead. The next stage may be developing the protocol for the main trial, a pilot study or the main trial itself*".

Fig 2. Definitions of pilot and feasibility studies used in on-line Delphi survey.

doi:10.1371/journal.pone.0150205.g002

version of the survey with a purposive sample of researchers active in the field of trials and pilot studies, and a workshop at the 2013 Society for Clinical Trials Conference in Boston, we further refined the definitions, checklists, survey introduction and added additional questions.

The first round of the main Delphi survey included: a description and explanation of our definitions of pilot and feasibility studies including examples (Figs 2 and 3); questions about participants' characteristics; 67 proposed items for the two checklists and questions about overall appropriateness of the guidelines for feasibility or pilot studies; and four questions related to the definitions of feasibility and pilot studies: *How appropriate do you think our definition for a pilot study conducted in preparation for an RCT is*? *How appropriate do you think our definition for a feasibility study conducted in preparation for an RCT is*? *How appropriate is the way we have distinguished between two different types of study conducted in preparation for an RCT*? *How appropriate are the labels 'pilot' and 'feasibility' for the two types of study we have*

Randomised pilot study:
Heazell AE *et al* [10] In this study the investigators wanted to assess whether a randomised controlled trial of the management of reduced fetal movement was feasible in relation to recruitment and retention, acceptability and adherence to protocol. They also wanted to confirm the prevalence of poor perinatal outcomes.

Non-randomised pilot study:
Colon HM *et al* [11] In this study investigators implemented an intervention to avoid the use of syringes and contamination of materials amongst injecting drug users. The intervention had four components and the investigators looked at the adoption of each component amongst a sample of 37 drug users recruited into the study. They also assessed whether the extent of blood residues had reduced sufficiently from baseline and post-intervention to indicate that this intervention merited further testing.

Feasibility study, not a pilot study:
Palmer AJ *et al* [12] In this study questionnaires were sent to surgeons and patients to determine their opinion about whether it would be feasible to conduct a randomised controlled trial comparing operative with non-operative treatment for femoroacetabular impingement surgery.

Fig 3. Examples of different types of pilot and feasibility study used in the on-line Delphi survey [10, 11, 12].

doi:10.1371/journal.pone.0150205.g003

distinguished? Participants were asked to rate their answers to the four questions on a nine-point scale from 'not at all appropriate' to 'completely appropriate'. There was also a space for open comments about the definitions. The second round included results from the first round and again asked for further comments about the definitions.

Participants for the main survey were identified as likely users of the checklist including trialists, methodologists, statisticians, funders and journal editors. Three hundred and seventy potential participants were approached by email from the project team or directly from Clinvivo. These were individuals identified based on personal networks, authors of relevant studies in the literature, members of the Canadian Institute of Health Research, Biostatistics section of Statistics Society of Canada, and the American Statistical Society. The International Society for Clinical Biostatistics and the Society for Clinical Trials kindly forwarded our email to their entire membership. There was a link within the email to the on-line questionnaire. Each round lasted three weeks and participants were sent one reminder a week before the closure of each survey. The survey took place between August and October 2013. Ethical approval was granted by the ScHARR research ethics committee at the University of Sheffield.

Developing a conceptual framework—Open meeting and research team meetings

The results of the Delphi survey pertaining to the definitions of feasibility and pilot studies were presented to an open meeting at the 2nd UK MRC Trials Methodology Conference in Edinburgh in November 2013 [13]. Attendees chose their preferred proposition from four propositions regarding the definitions, based variously on our original definitions, the NIHR and MRC views of pilot and feasibility studies and different views expressed in the Delphi survey. At a subsequent two-day research team meeting we collated the findings from the Delphi survey and the open meeting, and considered definitions of piloting and feasiblity outside the health research context found from on-line searches using the terms 'pilot definition', 'feasibility definition', 'pilot study definition' and 'feasibility study definition' in Google. We expected all searches to give a very large number of hits and examined the first two pages of hits only from each search. From this, we developed a conceptual framework reflecting consensus about the definitions, types and roles of feasibility and pilot studies conducted in preparation for an RCT evaluating the effect of an intervention or therapy. To ensure we incorporated the views of all researchers likely to be conducting pilot/feasiblity studies, two qualitative researchers joined the second day of the meeting which focused on agreeing this framework. Throughout this process we continually referred back to examples that we had identified to check that our emerging definitions were workable.

Validating the conceptual framework—systematic review

To validate the proposed conceptual framework, we identified a selection of recently reported studies that fitted our definition of pilot and feasibility studies, and tested a number of hypotheses in relation to these studies. We expected that approximately 30 reports would be sufficient to test the hypotheses. We conducted a systematic review to identify studies that authors described as pilot or feasibility studies, by searching Medline via PubMed for studies that had the words 'pilot' or 'feasibility' in the title. To increase the likelihood that the studies would be those conducted in preparation for a randomised controlled trial of the effect of a therapy or intervention we limited our search to those that contained the word 'trial' in the title or abstract. For full details of the search strategy see S1 Fig.

To focus on current practice, we selected the 150 most recent studies from those identified by the electronic search. We did not exclude protocols since we were primarily interested in

identifying the way researchers characterised their study and any possible future study and the relationship between them; we expected investigators to describe these aspects of their studies in a similar way in protocols and reports of findings. Two research team members independently reviewed study abstracts to assess whether each study fitted our working definition of a pilot or feasibility study in preparation for an RCT evaluating the effect of an intervention or therapy. Where reviewers disagreed, studies were classed as 'possible inclusions' and disagreements resolved by discussion with referral to the full text of the paper as necessary. Given the difficulty of interpreting some reports and to ensure that all research team members agreed on inclusion, the whole team then reviewed relevant extracted sections of the papers provisionally agreed for inclusion. We recognised that abstracts of some studies might not include appropriate information, and therefore that our initial abstract review could have excluded some relevant studies; we explored the extent of this potential omission of studies by reviewing the full texts of a random sample of 30 studies from the original 150. Since our prime goal was to identify a manageable number of relevant studies in order to test our hypotheses rather than identify all possible relevant studies we did not include any additional studies as a result of this exploratory study.

We postulated that the following hypotheses would support our conceptual framework:

1. The words 'pilot' and 'feasibility' are both used in the literature to describe studies undertaken in preparation for an RCT evaluating the effect of an intervention or therapy
2. It is possible to identify a subset of studies within the literature that are RCTs conducted in preparation for a larger RCT which evaluates the effect of an intervention or therapy. Authors do not use the term 'pilot trial' consistently in relation to these studies.
3. Within the literature it is not possible to apply unique mutually exclusive definitions of pilot and feasibility studies in preparation for an RCT evaluating the effect of an intervention or therapy that are consistent with the way authors describe their studies.
4. Amongst feasibility studies in preparation for an RCT which evaluates the effect of an intervention or therapy it is possible to identify some studies that are not pilot studies as defined within our conceptual framework, but are studies that acquire information about the feasibility of applying an intervention in a future study.

In order to explore these hypotheses, we categorised included studies into three groups that tallied with our framework (see results for details): randomised pilot studies, non-randomised pilot studies, feasibility studies that are not pilot studies. We also extracted data on objectives, and the phrases that indicated that the studies were conducted in preparation for a subsequent RCT.

Validating the conceptual framework—Consensus meeting

We also took an explanation and visual representation of our framework to an international consensus meeting primarily designed to reach consensus on an extension of the 2010 CONSORT statement to randomised pilot studies. There were 19 invited participants with known expertise, experience, or interest in pilot and feasibility studies, including representatives of CONSORT, funders, journal editors, and those who had been involved in writing the NIHR definitions of pilot and feasibility studies and the MRC guidance on designing and evaluating complex interventions. Thus this was an ideal forum in which to discuss the framework also. This project was not concerned with any specific disease, and was methodological in design; no patients or public were involved.

Results

Developing a conceptual framework—Delphi study

Ninety-three individuals, including chief investigators, statisticians, trial managers, clinicians, research assistants and a funder, participated in the first round of the Delphi survey and 79 in the second round. Over 70% of participants in the first round felt that our definitions, the way we had distinguished between pilot and feasibility studies, and the labels 'pilot' and 'feasibility' were appropriate. However, these four items had some of the lowest appropriateness ratings in the survey and there were a large number of comments both in direct response to our four survey items related to appropriateness of definitions, and in open comment boxes elsewhere in the survey. Some of these comments are presented in Fig 4. Some participants commented favourably on the definitions we had drawn up (quote 1) but others were confused by them (quote 2). Several compared our definitions to the NIHR definitions pointing out the differences (quote 3) and suggesting this might make it particularly difficult for the research community to understand our definitions (quote 4). Some expressed their own views about the definitions (quote 5); largely these tallied with the NIHR definitions. Others noted that both the concept of feasibility and the word itself were often used in relation to studies which investigators referred to as pilot studies (quote 6). Others questioned whether it was practically and/or theoretically possible to make a distinction between pilot and feasibility studies (quote 6, quote 7), suggesting that the two terms are not mutually exclusive and that feasibility was more of an umbrella term for studies conducted prior to the main trial. Some participants felt that, using our definitions, feasibility studies would be less structured and more variable and therefore their quality would be less appropriately assessed via a checklist (quote 8). These responses regarding definitions mirrored what we had found in the user-testing of the Delphi survey, the Society for Clinical Trials workshop, and differences of opinion already apparent in the literature. In the second round of the survey there were few comments about definitions.

Quote 1: Definitions could be tweaked but basically fine; problem is many people are not aware of any differences, let alone subtle ones.

Quote 2: Distinctions between pilot/feasibility are confusing and ambiguous. MRC say 'pilot need not be scale model' of proposed trial and warns against idea of linear progression.

Quote 3: Your definition of pilot more helpful (than NIHRs) by differentiating studies where an intervention is delivered from other studies where it is not. Quote 3

Quote 4: If a checklist or formal definitions are developed then it may take some time to get the community to understand the difference between the terms, especially as differ from NIHR.

Quote 5: Need cutting definition – for me pilot study is 'mimic' of main trial, feasibility study covers anything else.

Quote 6: In your definition pilot studies still tell us about feasibility of recruiting participants etc. so are all 'feasibility studies'.

Quote 7: Not sure two terms are mutually exclusive.

Quote 8: Feasibility studies (under your definition) are likely to be highly variable and a detailed checklist will be less useful. Pilot studies are likely to be of a predictable structure and will benefit from a checklist.

Fig 4. Quotes from the on-line Delphi survey.

doi:10.1371/journal.pone.0150205.g004

Developing a conceptual framework—Open meeting and research team meetings

There was a wide range of participants in the open meeting, including senior quantitative and qualitative methodologists, and a funding body representative. The four propositions we devised to cover different views about definitions of pilot and feasibility studies are shown in Fig 5. Fourteen out of the fifteen attendees who voted on these propositions preferred propositions 3 or 4, based on comments from the Delphi survey and the MRC guidance on designing and evaluating complex interventions respectively. Neither of these propositions implied mutually exclusive definitions of pilot and feasibility studies.

Definitions of feasibility outside the health research context focus on the likelihood of being able to do something. For example, the Oxford on-line dictionary defines feasibility as: 'The state or degree of being easily or conveniently done' [14] and a feasibility study as: 'An assessment of the practicality of a proposed plan or method' [15]. Some definitions also suggest that a feasibility study should help with decision making, for example [16]: 'The feasibility study is an evaluation and analysis of the potential of a proposed project. It is based on extensive investigation and research to support the process of decision making'. Outside the health research context the word 'pilot' has several different meanings but definitions of pilot studies usually focus on an experiment, project or development undertaken in advance of a future wider experiment, project or development. For example the Oxford on-line dictionary describes a pilot study as: 'Done as an experiment or test before being introduced more widely' [17]. Several definitions carry with them ideas that the purpose of a pilot study is also to facilitate decision making, for example 'a small-scale experiment or set of observations undertaken to decide how and whether to launch a full-scale project' [18] and some definitions specifically mention feasibility, for example: 'a small scale preliminary study conducted in order to evaluate feasibility' [19].

In keeping with these definitions not directly related to the health research context, we agreed that feasiblity is a concept encapsulating ideas about whether it is possible to do something and that *a feasibility study asks whether something can be done*, *should we proceed with it*, *and if so*, *how*. While piloting is also concerned with whether something can be done and whether and how we should proceed with it, it has a further dimension; piloting is implementing something, or part of something, in a way you intend to do it in future to see whether it can be done in practice. We therefore agreed that *a pilot study is a study in which a future study or part of a future study*, *is conducted on a smaller scale to ask the question whether something can be done*, *should we proceed with it*, *and if so*, *how*. The corollary of these definitions is that all pilot studies are feasibility studies but not all feasibility studies are pilot studies. Within the context of RCTs, the focus of our research, the 'something' in the definitions can be replaced with 'a future RCT evaluating the effect of an intervention or therapy'. Studies that address the question of whether the RCT can be done, should we proceed with it and if so how, can then be classed as feasibility or pilot studies. Some of these studies may, of course, have other objectives but if they are mainly focusing on feasiblity of the future RCT we would include them as feasiblity studies. All three studies used as examples in our Delphi survey [10–12] satisfy the definition of a feasiblity study. However, a study by Piot *et al*, that we encountered while developing the Delphi study, does not. This study is described as a pilot trial in the abstract but the authors present only data on effectiveness and although they state that their results require confirmation in a larger study it is not clear that their pilot study was conducted in preparation for such a larger study [20]. On the other hand, Palmer *et al* 'performed a feasibility study to determine whether patient and surgeon opinion was permissive for a Randomised Controlled Trial (RCT) comparing operative with non-operative treatment for FAI [femoroacetabular impingement]'

Definition 1 (our original proposal)

A feasibility study for a randomised controlled trial answers the question "Can these potential parts of a trial be done?" It has a clearly defined objective or set of objectives. These objectives need to be met for the next stage of study development to go ahead. The next stage may be developing the protocol for the main trial, a pilot study or the main trial itself".
A pilot study for a randomised controlled trial focuses on the integrity of the study protocol for the main trial. It asks the question 'Will this protocol work, if not why not and how should it be changed?' It may therefore resemble the main trial in many respects, particularly in study design and conduct. However the aim is not to test the effectiveness of the intervention but to assess the feasibility of the protocol.
Proposal: Two separate checklists are needed

Definition 2 (NIHR glossary – definition agreed by the EME, PHR, HTA and RfPB programs) (5, 6)

Feasibility Studies are pieces of research done before a main study in order to answer the question "Can this study be done? They are used to estimate important parameters that are needed to design the main study. For instance:

- standard deviation of the outcome measure, which is needed in some cases to estimate sample size;
- willingness of participants to be randomised;
- willingness of clinicians to recruit participants;
- number of eligible patients, carers or other appropriate participants
- characteristics of the proposed outcome measure and in some cases feasibility studies might involve designing a suitable outcome measure;
- Follow-up rates, response rates to questionnaires, adherence/compliance rates, ICCs in cluster trials, etc.
- availability of data needed or the usefulness and limitations of a particular database
- time needed to collect and analyse data

Pilot studies are a smaller version of the main study used to test whether the components of the main study can all work together. It is focused on the processes of the main study, for example to ensure that recruitment, randomisation, treatment, and follow-up assessments all run smoothly. It resembles the main study in many respects, including an assessment of the primary outcome. In some cases, this will be the first phase of the substantive study and data from the pilot phase may contribute to the final analysis; this can be referred to as an internal pilot. Or, at the end of the pilot study, the data may be analysed and set aside, a so-called external pilot.
Proposal: Two separate checklists are needed

Definition 3 (based on comments from Delphi)

The terms 'feasibility' and 'pilot' are not mutually exclusive. They are used interchangeably in the literature and it would be confusing to try and separate them out into two artificial sets of definitions.
A feasibility study is a multi-component study which may include qualitative research and 'developmental' type questions. It has a less rigid structure and there could be more variability in its conduct. An iterative approach to data collection in both quantitative and qualitative components may be taken, and the feasibility of actually delivering the intervention as well as actually running the trial may be a focus. There might be unexpected findings, and fundamental changes to the protocol. The results of a feasibility study might not be published but just be mentioned in the final main study manuscript.
A feasibility study may be followed by a pilot trial. Efficacy work and effectiveness work may both require feasibility and/or pilot stages.
Pilot trials have as their primary objective the assessment of factors that will determine whether an RCT can or should be done. Phase 1 and Phase 2 trials can be considered as pilot trials for the Phase 3 trial.
Proposal 3a: One checklist is needed to cover all pre-trial work

Proposal 3b: One checklist is needed for pilot trials only (it is impossible for a checklist to cover all desirable features of pilot/feasibility studies)

Definition 4 (based on MRC guidelines for complex interventions (2)

'The feasibility and piloting stage includes testing procedures for their acceptability, estimating the likely rates of recruitment and retention of subjects, and the calculation of appropriate sample sizes.
A pilot study need not be a 'scale model' of the planned main stage evaluation, but should address the main uncertainties that have been identified in the development work'
(www.mrc.ac.uk/complexinterventionsguidance)
The implication of the above is that this is a single category and we should use the same guidelines for all of them, with allowances for omitting sections that are not applicable to the particular preliminary study.
Proposal: One checklist is needed to cover all pre-trial work

Fig 5. Four propositions presented at Edinburgh open meeting.

doi:10.1371/journal.pone.0150205.g005

[12]. Heazell *et al* describe the aim of their randomised study as 'to address whether a randomised controlled trial (RCT) of the management of RFM [reduced fetal movement] was feasible' [10]. Their study was piloting many of the aspects they hoped to implement in a larger trial of RFM, thus making this also a pilot study, whereas the study conducted by Palmer *et al*, which comprised a questionnare to clinicians and seeking patient opinion, is not a pilot study but is a feasibility study.

Within our framework, some important studies conducted in advance of a future RCT to evaluate the effect of a therapy or intervention are not feasibility studies. For example, a systematic review, usually an essential pre-requisite for such an RCT, normally addresses whether the future RCT is *necessary* or *desirable*, not whether it is *feasible*. To reflect this, we developed a comprehensive diagrammatical representation of our framework for studies conducted in preparation for an RCT which, for completeness, includes, on the left hand side, early studies that are not pilot and feasibility studies, such as systematic reviews and, along the bottom, details of existing or planned reporting guidelines for different types of study (S2 Fig).

Validating the conceptual framework—Systematic review

From the 150 most recent studies identified by our electronic search, we identified 27 eligible reports (Fig 6). In keeping with our working definition of a pilot or feasibility study, to be included the reports had to show evidence that investigators were addressing at least some feasibility objectives and that the study was in preparation for a future RCT evaluating the effect of an intervention. Ideally we would have stipulated that the primary objective of the study should be a feasibility objective but, given the nature of the reporting of most of these studies, we felt this would be too restrictive.

The 27 studies are reported in Table 1 and results relating to terminology that authors used summarised in Table 2. Results in Table 2 support our first hypothesis that the words 'pilot' and 'feasibility' are both used in the literature to describe studies undertaken in preparation for a randomised controlled trial of effectiveness; 63% (17/27) used both terms somewhere in the title or abstract. The table also supports our second hypothesis that amongst the subset of feasibility studies in preparation for an RCT that are themselves RCTs, authors do not use the term 'pilot trial' consistently in relation to these studies; of the 18 randomised studies only eight contained the words 'pilot' and 'trial' in the title. Our third hypothesis, namely that it is not possible to apply unique mutually exclusive definitions of pilot and feasibility studies in preparation for an RCT that are consistent with the way authors describe their studies, is supported by the characteristics of studies presented in Table 1 and summarised in Table 2. We could find no design or other features (such as randomisation or presence of a control group) that distinguished between those that investigators called feasibility studies and those that they called pilot studies. However, the fourth hypothesis, that amongst studies in preparation for an RCT evaluating the effect of an intervention or therapy it is possible to identify some studies that explore the feasibility of a certain intervention or acquire related information about the feasibility of applying an intervention in a future study but are not pilot studies, was not supported; we identified no such studies amongst those reported in Table 1. Nevertheless, we had identified two prior to carrying out the review [10, 15].

Out of our exploratory sample of 30 study reports for which we reviewed full texts rather than only titles and abstracts, we identified 10 that could be classed as pilot or feasibility studies using our framework. We had already identified four of these in our sample reported in Table 1, but had failed to identify the other six. As expected, this was because key information to identify them as pilot or feasiblity studies such as the fact that they were in preparation for a larger RCT, or that the main objectives were to do with feasiblity were not included in the

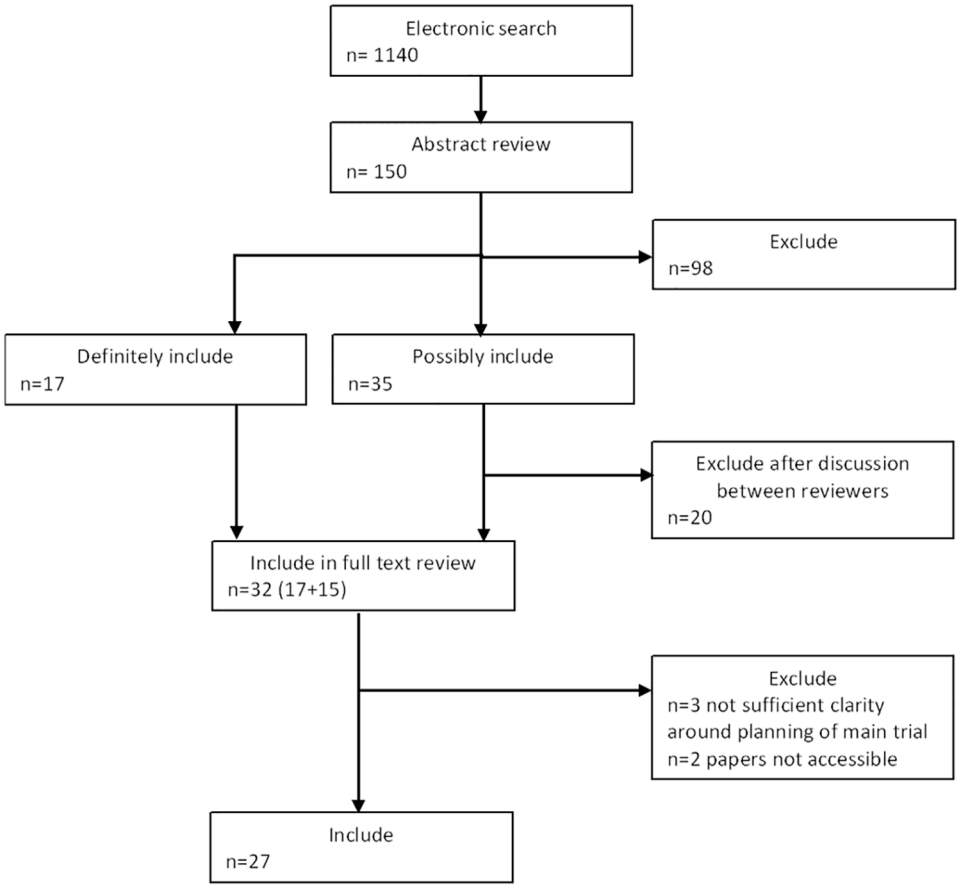


Fig 6. Flow chart showing identification of empirical pilot and feasibility studies.

doi:10.1371/journal.pone.0150205.g006

abstract. Thus our assumption that an initial screen using only abstracts resulted in the omission of some pilot and feasiblity studies was correct.

Validating the conceptual framework—Consensus meeting

International consensus meeting participants agreed with the general tenets of our conceptual framework including the ideas that all pilot studies are feasibility studies but that some feasibility studies are not pilot studies. They suggested that any definitive diagrammatic representation should more strongly reflect non-linearity in the ordering of feasibility studies. As a result of their input we produced a new, simplified, diagrammatical representation of the framework (Fig 7) which focuses on the key elements represented inside an oval shape on our original diagram, omits the wider context outside this shape, and highlights some features, including the non-linearity, more clearly.

The finalised framework

Fig 7 represents the framework. The figure indicates that where there is uncertainty about future RCT feasibility, a feasibility study is appropriate. Feasibility is thus an overarching concept within which we distinguish between three distinct types of study. *Randomised pilot studies* are those studies in which the future RCT, or parts of it, including the randomisation of

Table 1. Categorisation and description of the 27 pilot/feasibility studies identified in our systematic review.

	Pilot in title or abstract	Feasibility in title or abstract	Objectives	Phrase indicating that this is a pilot/ feasibility study in preparation for a future definitive trial	Trial in title or abstract
Randomised pilot/feasibility studies					
Allen [21]	-	Title and abstract	The study purpose was to assess the feasibility of recruiting pregnant adolescents into a randomised controlled trial, in order to inform the design of an adequately powered trial which could test the effect of caseload midwifery on preterm birth for pregnant adolescents.	. . .. in order to inform the design of an adequately powered trial which could test the effect of caseload midwifery on preterm birth for pregnant adolescents.	Title and abstract
Boogerd [22]	-	Title and abstract	To evaluate the feasibility of an online interactive treatment environment for adolescents with type 1 diabetes, called Sugarsquare, to supplement usual care	Results are promising and next steps are a full-scale randomised controlled trial and subsequent implementation in daily care.	Abstract
Buse [23]	Title and abstract	Abstract	We undertook a pilot trial to determine the feasibility of a trial comparing accelerated care (i.e., rapid medical clearance and surgery) and standard care among patients with a hip fracture.	These results show the feasibility of a trial comparing accelerated and standard care among patients with hip fracture and support a definitive trial. . ... Finally, this pilot trial identified design issues that we were able to overcome through protocol amendments.	Title and abstract
Clark [24]	Title and abstract	Abstract	The primary aim of this pilot trial was to assess the feasibility and safety of asking adults with stage 3 CKD to follow the above hydration intervention.	Prior to initiating a larger randomised controlled trial (RCT), we examined the safety and feasibility of asking adults with chronic kidney disease (CKD) to increase their water intake.	Title and abstract
Crawley [25]	Abstract	Title and abstract	Integrated qualitative methodology was used to explore the feasibility and acceptability of the recruitment, randomisation and interventions.	As the aim of this study was to assess the feasibility of a future definitive trial, we did not undertake a formal sample size calculation.	Title and abstract
Goodall [26]	Title and abstract	-	To this end, our trial had three objectives: piloting of trial processes; a quantitative measurement of changes in heart healthy behaviours with an economic evaluation (results published) and a qualitative evaluation of LHTs training and intervention delivery, implementation and acceptability (results to be reported elsewhere).	Our pilot explored feasibility of an LHT intervention before embarking on a full RCT.	Title and abstract
Higgins [27]	Title	Abstract	Evaluate the feasibility of a randomized controlled trial aimed at determining the efficacy of rTMS as an adjunct to task-oriented therapy in facilitating restoration of arm function after stroke.	Evaluate the feasibility of a randomized controlled trial. . ..	Title and abstract
Holt [28]	Title and abstract	Abstract	We plan a large, definitive, primary-care-based trial to determine efficacy and safety in patients with rotator cuff tendinopathy, and conducted a pilot trial to explore feasibility.	The lessons learned from this pilot will usefully inform the design of a large, definitive efficacy trial in primary care.	Title and abstract
Hurt [29]	Abstract	Title and abstract	This trial will assess the feasibility and inform the design of a large, UK-wide, clinical trial of a change to the NICE guidelines for urgent referral for chest X-ray for suspected lung cancer.	. . ...and inform the design of a large, UK-wide, clinical trial. . ..	Title and abstract

(*Continued*)

Table 1. (*Continued*)

	Pilot in title or abstract	Feasibility in title or abstract	Objectives	Phrase indicating that this is a pilot/ feasibility study in preparation for a future definitive trial	Trial in title or abstract
Lakes [30]	Title and abstract	Title	The objective of this pilot study was to evaluate Taekwondo implemented in public middle school physical education (PE). . ... Together, academic and community partners developed the current pilot study to address the feasibility and acceptability of implementing Taekwondo into PE in a public, low-income middle school as well as to investigate the effects of Taekwondo	Therefore, this pilot study lacked sufficient power to measure effects with statistical significance, but was expected to be sufficient to note trends in improvements that could be studied in a subsequent larger study.	Abstract
Lee [31]	Title and abstract	Abstract	Here, we examine the feasibility of the BCI system with a new game that incorporates memory training in improving memory and attention in a pilot sample of healthy elderly.	Obtain an estimate of efficacy in improving memory and attention in healthy elderly participants to determine whether the study should proceed to a phase III trial.	Abstract
McKenna [32]	-	Title and abstract	The aim of this randomized controlled trial was to evaluate the feasibility of delivering the Bridges stroke self-management program in addition to usual stroke rehabilitation compared with usual rehabilitation only.	A range of outcome measures were used to test their feasibility and explore whether they would be meaningful to use in a fully powered trial. . ... . .it would be advisable in future trials to keep more detailed records regarding the time spent on each component.	Title and abstract
Powell [33]	Title and abstract	Title and abstract	This article presents the findings of a pilot economic evaluation study running alongside the Bristol Girls Dance Project (BGDP) feasibility study.	. . .using a pilot economic evaluation to inform design of a full trial	Title and abstract
Saez [34]	Title and abstract	Abstract	In this work, we present the results of a randomized pilot study to evaluate the feasibility and to define the potential value for clinical practice of Curiam BT,. . .	We used these results as a baseline for the estimation of the total number of cases required to obtain statistical significant difference ($\alpha = .05$) in a larger RCT for the discrimination of tumour grades (Q2).	Abstract
Safdar [35]	Title and abstract	-	We aim to develop and evaluate a behavioural intervention 'Smoke Free Homes' (SFH) for TB patients that encourages them to negotiate a smoke free environment within their homes.	This is a pilot individual randomised controlled trial of SFH that will inform the design of a future definitive trial.	Title and abstract
Schoultz [36]	Title	Abstract	The aim of this study is to obtain the information required to design a full scale randomised controlled trial (RCT) that will examine the effectiveness of MBCT in improving quality of life for IBD patients.	The data will inform the estimate for recruitment rates for a full trial	Title and abstract
Siriwardhana [37]	Title and abstract	Abstract	The proposed pilot study aims to explore the feasibility of integrating mental health care into primary care by providing training to primary care practitioners serving displaced populations, in order to improve identification, treatment,and referral of patients with common mental disorders via the World Health Organization Mental Health Gap Action	Results will be used to formulate sample size calculation for a larger intervention.	Abstract
Wolf [38]	Title and abstract	-	The aim of the work presented here is to reduce the number of falls on a geriatric ward by monitoring patients more closely. To achieve this goal, a bed-exit alarm that reliably detects an attempt to get up has been constructed.	There are plans for a larger multicenter clinical trial to fortify these results. However, to be able to equip clinics on a larger scale and reach more patients, some modifications to the hardware are needed.	Abstract
Non-randomised pilot/feasibility studies					

(*Continued*)

Table 1. (*Continued*)

	Pilot in title or abstract	Feasibility in title or abstract	Objectives	Phrase indicating that this is a pilot/ feasibility study in preparation for a future definitive trial	Trial in title or abstract
Alers [39]	Title	-	A phase I clinical trial to investigate the efficacy of maternal oral melatonin administration in women with a pregnancy complicated by fetal growth restriction	If this trial is successful, the results will be used to inform future randomised controlled trials.	Title and abstract
Carlesso [40]	Title and abstract	Title and abstract	To pilot and determine the feasibility of estimating adverse events in patients with neck pain treated with cervical manipulation/mobilization by Canadian orthopaedic manual physiotherapists (OMPTs) using an online data-collection system to provide estimates.	. . ...to provide estimates for a future larger multi-centre international study.	Abstract
Collado [41]	Title	-	to evaluate BATD, an idiographic intervention, employing the rationale that BATD provides a flexible and easily-tailored treatment framework able to address the individual and psychological needs of depressed Latinos.	The study's positive outcomes suggest that a Stage II randomized clinical trial is a logical next step.	Abstract
Galantino [42]	Abstract	Title and abstract	This study aimed to determine the feasibility of tai chi to improve well-being for women experiencing AI-associated arthralgias (AIAAs).	The sample size of this pilot study was not intended to provide an efficacy analysis but rather to obtain an estimate of the effect size and variance necessary to plan a definitive study to test and refine individual components of the tai chi protocol for AIAA and measurement tools.	Abstract
Garcia [43]	Title and abstract	Abstract	Prior to implementing a large randomized trial at our institution, we investigated the feasibility, safety, and initial efficacy of acupuncture for uncontrolled pain among cancer patients.	Prior to implementing a large randomized trial at our institution.	Abstract
Hu [44]	-	Title and abstract	To determine the feasibility of all aspects of a pragmatic observational study designed: (1) to evaluate the effectiveness and cost effectiveness of integrated treatments for MSDs in an integrated NHS hospital in the UK; (2) to determine the acceptability of the study design and research process to patients; (3) to explore patients' expectation and experience of receiving integrated treatments.	It will inform the design of a future trial including recruitment, retention, suitability of the outcome measures and patients' experiences.	Abstract
Misumi [45]	-	Title and abstract	We conducted a feasibility study to evaluate the safety and efficacy of carboplatin plus irinotecan in preparation for a planned Phase III study.	Based on these results, a Phase II/III trial comparing carboplatin plus etoposide with carboplatin plus irinotecan for elderly patients with extensive disease small-cell lung cancer is being planned by the Japan Clinical Oncology Group.	Abstract
Penn [46]	Title and abstract	Title and abstract	. . .aimed to assess the feasibility, acceptability and outcomes at a 12-month follow-up of a behavioural intervention for adults at risk of T2D.	Feasibility and acceptability of this novel intervention were assessed in preparation for a definitive effectiveness trial.	Abstract
Pompeu [47]	Title	Title and abstract	To assess the feasibility, safety and outcomes of playing Microsoft Kinect AdventuresTM for people with Parkinson's disease in order to guide the design of a randomised clinical trial.	. . .. in order to guide the design of a randomised clinical trial.	Abstract

doi:10.1371/journal.pone.0150205.t001

Table 2. Summary of terms used in 27 pilot/feasibility studies.

Use of the terms pilot and feasibility in the title and abstract	All included studies	Randomised studies	Non-randomised studies	Randomised studies with trial in the title
Pilot in title, no mention of feasibility in title or abstract	5	3	2	2
Feasibility in title, no mention of pilot in title or abstract	5	3	2	2
Both terms in title	5	2	3	1
Pilot in title, feasibility in abstract only	9	8	1	5
Feasibility in title, pilot in abstract only	3	2	1	2
Total	27	18	9	12

doi:10.1371/journal.pone.0150205.t002

participants, is conducted on a smaller scale (piloted) to see if it can be done. Thus randomised pilot studies can include studies that for the most part reflect the design of a future definitive trial but, if necessary due to remaining uncertainty, may involve trying out alternative strategies, for example, collecting an outcome variable via telephone for some participants and online for others. Within the framework randomised pilot studies could also legitimately be called randomised feasibility studies. Two-thirds of the studies presented in Table 1 are of this type.

Non-randomised pilot studies are similar to randomised pilot studies; they are studies in which all or part of the intervention to be evaluated and other processes to be undertaken in a future trial is/are carried out (piloted) but without randomisation of participants. These could also legitimately be called by the umbrella term, feasibility study. These studies cover a wide range from those that are very similar to randomised pilot studies except that the intervention and control groups have not been randomised, to those in which only the intervention, and no other trial processes, are piloted. One-third of studies presented in Table 1 are of this type.

Feasibility studies that are not pilot studies are those in which investigators attempt to answer a question about whether some element of the future trial can be done but do not implement the intervention to be evaluated or other processes to be undertaken in a future trial, though they may be addressing intervention development in some way. Such studies are rarer than the other types of feasibility study and, in fact, none of the studies in Table 1 were of this type. Nevertheless, we include these studies within the framework because they do exist; the Palmer study [15] in which surgeons and patients were asked about the feasibility of randomisation is one such example. Other examples might be interviews to ascertain the acceptability of an intervention, or questionnaires to assess the types of outcomes participants might think important. Within the framework these studies can be called feasibility studies but cannot be called pilot studies since no part of the future randomised controlled trial is being conducted on a smaller scale.

Investigators may conduct a number of studies to assess feasibility of an RCT to test the effect of any intervention or therapy. While it may be most common to carry out what we have referred to as *feasibility studies that are not pilot studies* before *non-randomised pilot studies*, and *non-randomised pilot studies* prior to *randomised pilot studies*, the process of feasibility work is not necessarily linear and such studies can in fact be conducted in any order. For completeness the diagram indicates the location of internal pilot studies.

Discussion

There are diverse views about the definitions of pilot and feasibility studies within the research community. We reached consensus over a conceptual framework for the definitions of these studies in which feasibility is an overarching concept for studies assessing whether a future

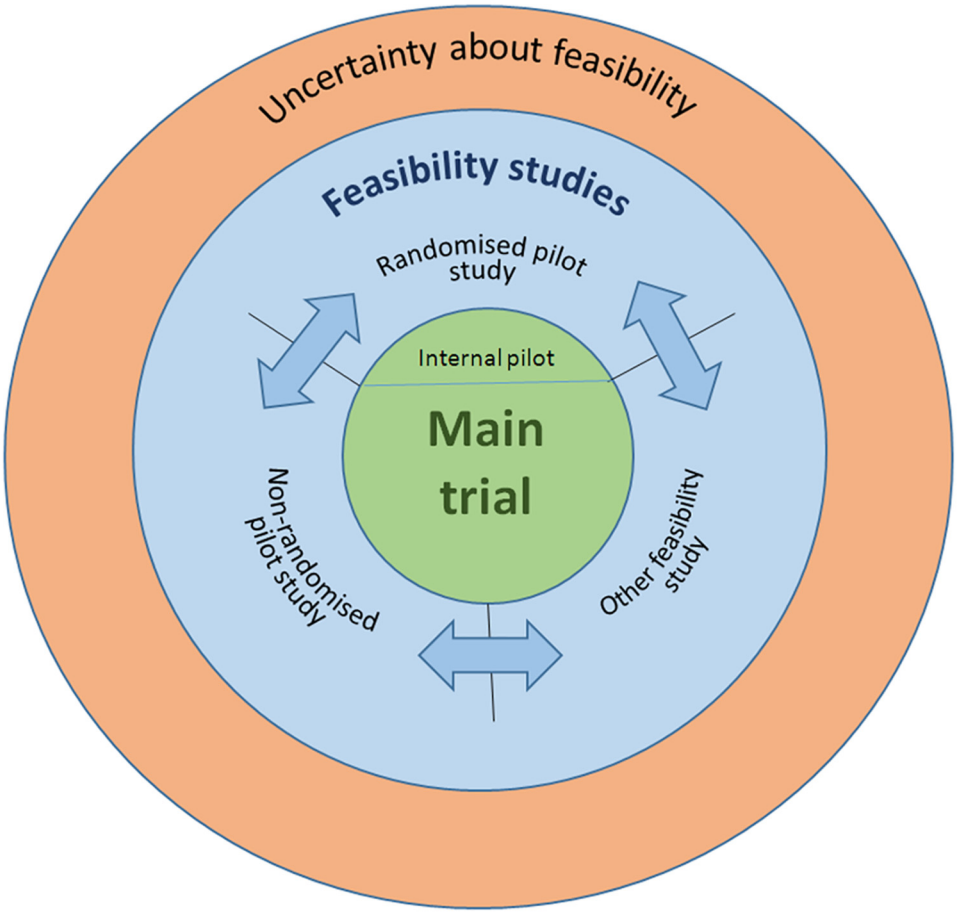


Fig 7. Conceptual framework.

doi:10.1371/journal.pone.0150205.g007

study, project or development can be done. For studies conducted in preparation for a RCT assessing the effect of a therapy or intervention, three distinct types of study come under the umbrella of feasibility studies: randomised pilot studies, non-randomised pilot studies, feasibility studies that are not pilot studies. Thus pilot studies are a subset of feasibility studies. A review of the literature confirmed that it is not possible to apply mutually exclusive definitions of pilot and feasibility studies in preparation for such an RCT that are consistent with the way authors describe their studies. For example Lee *et al* [31], Boogerd *et al* [22] and Wolf *et al* [38] all describe randomised studies exploring the feasibility of introducing new systems (brain computer interface memory training game, on-line interactive treatment environment, bed-exit alarm respectively) but Lee *et al* describe their study as a 'A Randomized Control Pilot Study', with the word 'feasibility' used in the abstract and text, while the study by Boogerd *et al.* is titled 'Teaming up: feasibility of an online treatment environment for adolescents with type 1 diabetes', and Wolf at al describe their study as a pilot study without using the word 'feasibility'.

Our re-evaluation of the definitions of pilot and feasibility studies was conducted over a period of time with input via a variety of media by multi-disciplinary and international researchers, publishers, editors and funders. It was to some extent a by-product of our work developing reporting guidelines for such studies. Nevertheless, we were able to gather a wide range of expert views, and the iterative nature of the development of our thinking has been an

important part of obtaining consensus. Other parallel developments, including the recent establishment of the new Pilot and Feasibility Studies journal [48], suggest that our work is, indeed, timely. We encountered several difficulties in reviewing empirical study reports. Firstly, it was sometimes hard to assess whether studies were planned in preparation for an RCT or whether the authors were conducting a small study and simply commenting on the fact that a larger RCT would be useful. Secondly, objectives were sometimes unclear, and/or effectiveness objectives were often emphasised in spite of recommendations that pilot and feasibility studies should not be focusing on effectiveness [1, 4]. In identifying relevant studies we erred on the side of inclusiveness, acknowledging that getting these studies published is not easy and that there are, as yet, no definitive reporting guidelines for investigators to follow. Lastly, our electronic search was unable to identify any feasibility studies that were not pilot studies according to our definitions. Subsequent discussion with qualitative researchers suggested that this is because such studies are often not described as feasibility studies in the title or abstract.

Our framework is compatible with the UK MRC guidance on complex interventions which suggests a 'feasibility and piloting' phase as part of the work to design and evaluate such interventions without any explicit distinction between pilot and feasibility studies. In addition, although our framework has a different underlying principle from that adopted by UK NIHR, the NIHR definition of a pilot study is not far from the subset of studies we have described as randomised pilot studies. Although there appears to be increasing interest in pilot and feasibility studies, as far as we are aware no other funding bodies specifically address the nature of such studies. The National Institute for Health in the USA does, however, routinely require published pilot studies before considering funding applications for certain streams, and the Canadian Institutes of Health Research routinely have calls for pilot or feasibility studies in different clinical areas to gather evidence necessary to determine the viability of new research directions determined by their strategic funding plans. These approaches highlight the need for clarity regarding what constitutes a pilot study.

There are several previous reviews of empirical pilot and feasibility studies [1, 4, 7]. In the most recent, reviewing studies published between 2000 and 2009 [7], the authors identified a large number of studies, described similar difficulty in identifying whether a larger study was actually being planned, and similar lack of consistency in the way the terms 'pilot' and 'feasibility' are used. Nevertheless, in methodological work, many researchers have adopted fairly rigid definitions of pilot and feasibility studies. For example, Bugge *et al* in developing the ADEPT framework refer to the NIHR definitions and suggest that feasibility studies ask questions about 'whether the study can be done' while pilot trials are '(a miniature version of the main trial), which aim to test aspects of study design and processes for the implementation of a larger main trial in the future' [49]. Although not explicitly stated, the text seems to suggest that pilot and feasibility studies are mutually exclusive. Our work indicates that this is neither necessary nor desirable. There is, however, general agreement in the literature about the purpose of pilot and feasibility studies. For example, pilot trials are 'to provide sufficient assurance to enable a larger definitive trial to be undertaken' [50], and pilot studies are 'designed to test the performance characteristics and capabilities of study designs, measures, procedures, recruitment criteria, and operational strategies that are under consideration for use in a subsequent, often larger, study' [51], and 'play a pivotal role in the planning of large-scale and often expensive investigations' [52]. Within our framework we define all studies aiming to assess whether a future RCT is doable as 'feasibility studies'. Some might argue that the focus of their study in preparation for a future RCT is acceptability rather than feasibility, and indeed, in other frameworks, such as the RE-AIM framework [53], feasibility and acceptability are seen as two different concepts. However, it is perfectly possible to explore the acceptability of an intervention, of a data collection process or of randomisation in order to determine the *feasibility* of a putative larger RCT. Thus

the use of the term 'feasibility study' for a study in preparation for a future RCT is not incompatible with the exploration of issues other than feasibility within the study itself.

There are numerous previous studies in which the investigators review the literature and seek the counsel of experts to develop definitions and clarify terminology. Most of these relate to clinical or physiological definitions [54–56]. A few explorations of definitions relate to concepts such as quality of life [57]. Implicit in much of this work is that from time to time definitions need rethinking as knowledge and practice moves on. From an etymological point of view this makes sense. In fact, the use of the word 'pilot' to mean something that is a prototype of something else only appears to emerge in the middle of the twentieth century and the first use of the word in relation to research design that we could find was in 1947—a pilot survey [58]. Thus we do not have to look very far back to see changes in the use of one of the words we have been dealing with in developing our conceptual framework. We hope what we are proposing here is helpful in the early twenty-first century to clarify the use of the words 'pilot' and 'feasibility' in a health research context.

We suggest that researchers view feasibility as an overarching concept, with all studies done in preparation for a main study open to being called feasibility studies, and with pilot studies as a subset of feasibility studies. All such studies should be labelled 'pilot' and/or 'feasibility' as appropriate, preferably in the title of a report, but if not certainly in the abstract. This recommendation applies to all studies that contribute to an assessment of the feasibility of an RCT evaluating the effect of an intervention. Using either of the terms in the title will be most helpful for those conducting future electronic searches. However, we recognise that for qualitative studies, authors may find it convenient to use the terms in the abstract rather than the title. Authors also need to describe objectives and methods well, reporting clearly if their study is in preparation for a future RCT to evaluate the effect of an intervention or therapy.

Though the focus of this work was on the definitions of pilot and feasibility studies and extensive recommendations for the conduct of these studies is outside its scope, we suggest that in choosing what type of feasibility study to conduct investigators should pay close attention to the major uncertainties that exist in relation to trial or intervention. A randomised pilot study may not be necessary to address these; in some cases it may not even be necessary to implement an intervention at all. Similarly, funders should look for a justification for the type of feasibility study that investigators propose. We have has also highlighted the need for better reporting of these studies. The CONSORT extension for randomised pilot studies that our group has developed are important in helping to address this need and will be reported separately. Nevertheless, further work will be necessary to extend or adapt these reporting guidelines for use for non-randomised pilot studies and for feasibility studies that are not pilot studies. There is also more work to be done in developing good practice guidance for the conduct of pilot and feasibility studies.

Supporting Information

S1 Fig. Search strategy to identify studies that authors described as pilot or feasibility studies.
(DOCX)

S2 Fig. Initial comprehensive diagrammatic representation of framework.
(DOCX)

Acknowledgments

We thank Alicia O'Cathain and Pat Hoddinot for discussions about the reporting of qualitative studies, and consensus participants for their views on our developing framework. Claire

Coleman was funded by a National Institute for Health Research (NIHR) Research Methods Fellowship. This article presents independent research funded by the NIHR. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

Author Contributions

Conceived and designed the experiments: SE GL MC LT SH CB. Performed the experiments: SE GL MC LT SH CB CC. Analyzed the data: SE GL MC LT SH CB CC. Contributed reagents/materials/analysis tools: SE GL MC LT SH CB. Wrote the paper: SE GL MC LT SH CB CC.

References

1. Lancaster GA, Dodd S, Williamson PR. Design and analysis of pilot studies: recommendations for good practice. Journal of evaluation in clinical practice. 2004; 10(2):307–12. Epub 2004/06/11. doi: 10.1111/j.2002.384.doc.x PMID: 15189396.
2. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M. Developing and evaluating complex interventions: the new Medical Research Council guidance. British Medical Journal. 2008; 337.
3. Thabane L, Ma J, Chu R, Cheng J, Ismaila A, Rios LP, et al. A tutorial on pilot studies: the what, why and how. BMC medical research methodology. 2010; 10:1. Epub 2010/01/08. doi: 10.1186/1471-2288-10-1 PMID: 20053272; PubMed Central PMCID: PMCPmc2824145.
4. Arain M, Campbell MJ, Cooper CL, Lancaster GA. What is a pilot or feasibility study? A review of current practice and editorial policy. BMC medical research methodology. 2010; 10:67. Epub 2010/07/20. doi: 10.1186/1471-2288-10-67 PMID: 20637084; PubMed Central PMCID: PMCPmc2912920.
5. National Institute for Health Research. NIHR Evaluation, Trials and Studies | Glossary 2015. Available: http://www.nets.nihr.ac.uk/glossary/feasibility-studies. Accessed 2015 Mar 17.
6. National Institute for Health Research. NIHR Evaluation, Trials and Studies | Pilot studies 2015. Available: http://www.nets.nihr.ac.uk/glossary/pilot-studies. Accessed 2015 Mar 17.
7. Shanyinde M, Pickering RM, Weatherall M. Questions asked and answered in pilot and feasibility randomized controlled trials. BMC medical research methodology. 2011; 11:117. Epub 2011/08/19. doi: 10.1186/1471-2288-11-117 PMID: 21846349; PubMed Central PMCID: PMCPmc3170294.
8. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. Annals of Internal Medicine. 2010; 152(11):726–32. Epub 2010/03/26. doi: 10.7326/0003-4819-152-11-201006010-00232 PMID: 20335313.
9. CLINVIVO. Clinvivo Limited 2015 [cited 2015 9 April]. Available: http://www.clinvivo.com/.
10. Heazell AE, Bernatavicius G, Roberts SA, Garrod A, Whitworth MK, Johnstone ED, et al. A randomised controlled trial comparing standard or intensive management of reduced fetal movements after 36 weeks gestation—a feasibility study. BMC pregnancy and childbirth. 2013; 13:95. Epub 2013/04/18. doi: 10.1186/1471-2393-13-95 PMID: 23590451; PubMed Central PMCID: PMCPmc3640967.
11. Colon HM, Finlinson HA, Negron J, Sosa I, Rios-Olivares E, Robles RR. Pilot trial of an intervention aimed at modifying drug preparation practices among injection drug users in Puerto Rico. AIDS and behavior. 2009; 13(3):523–31. Epub 2009/03/25. doi: 10.1007/s10461-009-9540-3 PMID: 19308722.
12. Palmer AJ, Thomas GE, Pollard TC, Rombach I, Taylor A, Arden N, et al. The feasibility of performing a randomised controlled trial for femoroacetabular impingement surgery. Bone & joint research. 2013; 2 (2):33–40. Epub 2013/04/24. doi: 10.1302/2046-3758.22.2000137 PMID: 23610700; PubMed Central PMCID: PMCPmc3626218.
13. In Conference Ltd. 2nd Clinical Trials Methodology Conference | 18–19 November 2013 EICC, Edinburgh, Scotland 2013. Available: http://www.methodologyconference2013.org.uk/. Accessed 2015 Mar 17.
14. Oxford Dictionaries. Oxford Dictionaries | feasibility 2015. Available: http://www.oxforddictionaries.com/definition/english/feasibility. Accessed 2015 Mar 17.
15. Oxford Dictionaries. Oxford Dictionaries | feasibility study 2015. Available: http://www.oxforddictionaries.com/definition/english/feasibility-study. Accessed 2015 Mar 17.
16. Wikipedia. Feasibility study 2015 [cited 2015 17 March]. Available: http://en.wikipedia.org/wiki/Feasibility_study.
17. Oxford Dictionaries. Oxford Dictionaries | pilot 2015. Available from: http://www.oxforddictionaries.com/definition/english/pilot. Accessed 2015 Mar 17.

18. Collins. Collins English Dictionary | pilot study 2015. Available from: http://www.collinsdictionary.com/dictionary/english/pilot-study. Accessed 2015 Mar 17.
19. Wikipedia. Pilot experiment 2015. Available: http://en.wikipedia.org/wiki/Pilot_experiment. Accessed 2015 Mar 17.
20. Piot C, Croisille P, Staat P, Thibault H, Rioufol G, Mewton N, et al. Effect of cyclosporine on reperfusion injury in acute myocardial infarction. The New England journal of medicine. 2008; 359(5):473–81. Epub 2008/08/02. doi: 10.1056/NEJMoa071142 PMID: 18669426.
21. Allen J, Stapleton H, Tracy S, Kildea S. Is a randomised controlled trial of a maternity care intervention for pregnant adolescents possible? An Australian feasibility study. BMC medical research methodology. 2013; 13:138. Epub 2013/11/15. doi: 10.1186/1471-2288-13-138 PMID: 24225138; PubMed Central PMCID: PMCPmc4226005.
22. Boogerd EA, Noordam C, Kremer JA, Prins JB, Verhaak CM. Teaming up: feasibility of an online treatment environment for adolescents with type 1 diabetes. Pediatric diabetes. 2014; 15(5):394–402. Epub 2013/12/20. doi: 10.1111/pedi.12103 PMID: 24350732.
23. Buse GL, Bhandari M, Sancheti P, Rocha S, Winemaker M, Adili A, et al. Accelerated care versus standard care among patients with hip fracture: the HIP ATTACK pilot trial. Canadian Medical Association Journal. 2013; 186(1):E52–E60. doi: 10.1503/cmaj.130901 PMID: 24246589
24. Clark WF, Sontrop JM, Huang S-H, Gallo K, Moist L, House AA, et al. The chronic kidney disease Water Intake Trial (WIT): results from the pilot randomised controlled trial. BMJ Open. 2013; 3(12): e003666. doi: 10.1136/bmjopen-2013-003666 PMID: 24362012
25. Crawley E, Mills N, Beasant L, Johnson D, Collin SM, Deans Z, et al. The feasibility and acceptability of conducting a trial of specialist medical care and the Lightning Process in children with chronic fatigue syndrome: feasibility randomized controlled trial (SMILE study). Trials. 2013; 14:415. Epub 2013/12/07. doi: 10.1186/1745-6215-14-415 PMID: 24304689; PubMed Central PMCID: PMCPmc4235039.
26. Goodall M, Barton GR, Bower P, Byrne P, Cade JE, Capewell S, et al. Food for thought: pilot randomized controlled trial of lay health trainers supporting dietary change to reduce cardiovascular disease in deprived communities. Journal of Public Health (Oxford, England) 2014; 36(4):635–43. Epub 2013/11/28. doi: 10.1093/pubmed/fdt112 PMID: 24277778.
27. Higgins J, Koski L, Xie H. Combining rTMS and task-oriented training in the rehabilitation of the arm after stroke: a pilot randomized controlled trial. Stroke Research and Treatment. 2013; 2013. doi: 10.1155/2013/539146
28. Holt TA, Mant D, Carr A, Gwilym S, Beard D, Toms C, et al. Corticosteroid injection for shoulder pain: single-blind randomized pilot trial in primary care. Trials. 2013; 14:425. Epub 2013/12/12. doi: 10.1186/1745-6215-14-425 PMID: 24325987; PubMed Central PMCID: PMCPmc3878869.
29. Hurt CN, Roberts K, Rogers TK, Griffiths GO, Hood K, Prout H, et al. A feasibility study examining the effect on lung cancer diagnosis of offering a chest X-ray to higher-risk patients with chest symptoms: protocol for a randomized controlled trial. Trials. 2013; 14:405. Epub 2013/11/28. doi: 10.1186/1745-6215-14-405 PMID: 24279296; PubMed Central PMCID: PMCPmc4222751.
30. Lakes KD, Bryars T, Sirisinahal S, Salim N, Arastoo S, Emmerson N, et al. The Healthy for Life Taekwondo Pilot Study: A Preliminary Evaluation of Effects on Executive Function and BMI, Feasibility, and Acceptability. Mental health and physical activity. 2013; 6(3):181–8. Epub 2014/02/25. doi: 10.1016/j.mhpa.2013.07.002 PMID: 24563664; PubMed Central PMCID: PMCPmc3927879.
31. Lee TS, Goh SJ, Quek SY, Phillips R, Guan C, Cheung YB, et al. A brain-computer interface based cognitive training system for healthy elderly: a randomized control pilot study for usability and preliminary efficacy. PloS one. 2013; 8(11):e79419. Epub 2013/11/22. doi: 10.1371/journal.pone.0079419 PMID: 24260218; PubMed Central PMCID: PMCPmc3832588.
32. McKenna S, Jones F, Glenfield P, Lennon S. Bridges self-management program for people with stroke in the community: A feasibility randomized controlled trial. International journal of stroke: official journal of the International Stroke Society. 2013. Epub 2013/11/22. doi: 10.1111/ijs.12195 PMID: 24256085.
33. Powell JE, Carroll FE, Sebire SJ, Haase AM, Jago R. Bristol girls dance project feasibility study: using a pilot economic evaluation to inform design of a full trial. BMJ Open. 2013; 3(12):e003726. doi: 10.1136/bmjopen-2013-003726 PMID: 24362013
34. Saez C, Marti-Bonmati L, Alberich-Bayarri A, Robles M, Garcia-Gomez JM. Randomized pilot study and qualitative evaluation of a clinical decision support system for brain tumour diagnosis based on SV (1)H MRS: evaluation as an additional information procedure for novice radiologists. Computers in biology and medicine. 2014; 45:26–33. Epub 2014/02/01. doi: 10.1016/j.compbiomed.2013.11.009 PMID: 24480160.
35. Safdar N, Zahid R, Shah S, Fatima R, Cameron I, Siddiqi K. Tuberculosis patients learning about second-hand smoke (TBLASS): results of a pilot randomised controlled trial. The international journal of

tuberculosis and lung disease: the official journal of the International Union against Tuberculosis and Lung Disease. 2015; 19(2):237–43. Epub 2015/01/13. doi: 10.5588/ijtld.14.0615 PMID: 25574925.

36. Schoultz M, Atherton IM, Hubbard G, Watson AJ. The use of mindfulness-based cognitive therapy for improving quality of life for inflammatory bowel disease patients: study protocol for a pilot randomised controlled trial with embedded process evaluation. Trials. 2013; 14:431. Epub 2013/12/18. doi: 10.1186/1745-6215-14-431 PMID: 24341333; PubMed Central PMCID: PMCPmc3878510.

37. Siriwardhana C, Adikari A, Van Bortel T, McCrone P, Sumathipala A. An intervention to improve mental health care for conflict-affected forced migrants in low-resource primary care settings: a WHO MhGAP-based pilot study in Sri Lanka (COM-GAP study). Trials. 2013; 14:423. Epub 2013/12/11. doi: 10.1186/1745-6215-14-423 PMID: 24321171; PubMed Central PMCID: PMCPmc3906999.

38. Wolf KH, Hetzer K, zu Schwabedissen HM, Wiese B, Marschollek M. Development and pilot study of a bed-exit alarm based on a body-worn accelerometer. Zeitschrift fur Gerontologie und Geriatrie. 2013; 46(8):727–33. Epub 2013/11/26. doi: 10.1007/s00391-013-0560-2 PMID: 24271253.

39. Alers NO, Jenkin G, Miller SL, Wallace EM. Antenatal melatonin as an antioxidant in human pregnancies complicated by fetal growth restriction—a phase I pilot clinical trial: study protocol. BMJ Open. 2013; 3(12):e004141. doi: 10.1136/bmjopen-2013-004141 PMID: 24366583

40. Carlesso LC, Macdermid JC, Santaguida PL, Thabane L. Determining adverse events in patients with neck pain receiving orthopaedic manual physiotherapy: a pilot and feasibility study. Physiotherapy Canada Physiotherapie Canada. 2013; 65(3):255–65. Epub 2014/01/10. doi: 10.3138/ptc.2012-28 PMID: 24403696; PubMed Central PMCID: PMCPmc3740991.

41. Collado A, Castillo SD, Maero F, Lejuez CW, Macpherson L. Pilot of the brief behavioral activation treatment for depression in latinos with limited english proficiency: preliminary evaluation of efficacy and acceptability. Behavior therapy. 2014; 45(1):102–15. Epub 2014/01/15. doi: 10.1016/j.beth.2013.10.001 PMID: 24411118; PubMed Central PMCID: PMCPmc4103902.

42. Galantino ML, Callens ML, Cardena GJ, Piela NL, Mao JJ. Tai chi for well-being of breast cancer survivors with aromatase inhibitor-associated arthralgias: a feasibility study. Alternative therapies in health and medicine. 2013; 19(6):38–44. Epub 2013/11/21. PMID: 24254037.

43. Garcia MK, Driver L, Haddad R, Lee R, Palmer JL, Wei Q, et al. Acupuncture for treatment of uncontrolled pain in cancer patients: a pragmatic pilot study. Integrative cancer therapies. 2014; 13(2):133–40. Epub 2013/11/28. doi: 10.1177/1534735413510558 PMID: 24282103.

44. Hu X, Hughes J, Fisher P, Lorenc A, Purtell R, Park AL, et al. A pragmatic observational feasibility study on integrated treatment for musculoskeletal disorders: Design and protocol. Chinese journal of integrative medicine. 2013. Epub 2013/12/18. doi: 10.1007/s11655-013-1557-9 PMID: 24338185.

45. Misumi Y, Nishio M, Takahashi T, Ohyanagi F, Horiike A, Murakami H, et al. A feasibility study of carboplatin plus irinotecan treatment for elderly patients with extensive disease small-cell lung cancer. Japanese journal of clinical oncology. 2014; 44(2):116–21. Epub 2013/12/18. doi: 10.1093/jjco/hyt195 PMID: 24338555.

46. Penn L, Ryan V, White M. Feasibility, acceptability and outcomes at a 12-month follow-up of a novel community-based intervention to prevent type 2 diabetes in adults at high risk: mixed methods pilot study. BMJ Open. 2013; 3(11):e003585. doi: 10.1136/bmjopen-2013-003585 PMID: 24227871

47. Pompeu JE, Arduini LA, Botelho AR, Fonseca MB, Pompeu SM, Torriani-Pasin C, et al. Feasibility, safety and outcomes of playing Kinect Adventures! for people with Parkinson's disease: a pilot study. Physiotherapy. 2014; 100(2):162–8. Epub 2014/04/08. doi: 10.1016/j.physio.2013.10.003 PMID: 24703891.

48. Lancaster GA. Pilot and feasibility studies come of age! Pilot and Feasibility Studies. 2015; 1(1):1.

49. Bugge C, Williams B, Hagen S, Logan J, Glazener C, Pringle S, et al. A process for Decision-making after Pilot and feasibility Trials (ADePT): development following a feasibility study of a complex intervention for pelvic organ prolapse. Trials. 2013; 14:353. Epub 2013/10/29. doi: 10.1186/1745-6215-14-353 PMID: 24160371; PubMed Central PMCID: PMCPmc3819659.

50. Lee EC, Whitehead AL, Jacques RM, Julious SA. The statistical interpretation of pilot trials: should significance thresholds be reconsidered? BMC medical research methodology. 2014; 14:41. Epub 2014/03/22. doi: 10.1186/1471-2288-14-41 PMID: 24650044; PubMed Central PMCID: PMCPmc3994566.

51. Moore CG, Carter RE, Nietert PJ, Stewart PW. Recommendations for planning pilot studies in clinical and translational research. Clinical and translational science. 2011; 4(5):332–7. Epub 2011/10/28. doi: 10.1111/j.1752-8062.2011.00347.x PMID: 22029804; PubMed Central PMCID: PMCPmc3203750.

52. Brooks D, Stratford P. Pilot studies and their suitability for publication in physiotherapy Canada. Physiotherapy Canada Physiotherapie Canada. 2009; 61(2):66–7. Epub 2010/03/02. doi: 10.3138/physio.61.2.66 PMID: 20190988; PubMed Central PMCID: PMCPmc2792235.

53. Glasgow RE, Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. American journal of public health. 1999; 89(9):1322–7. Epub 1999/09/04. PMID: 10474547; PubMed Central PMCID: PMCPmc1508772.

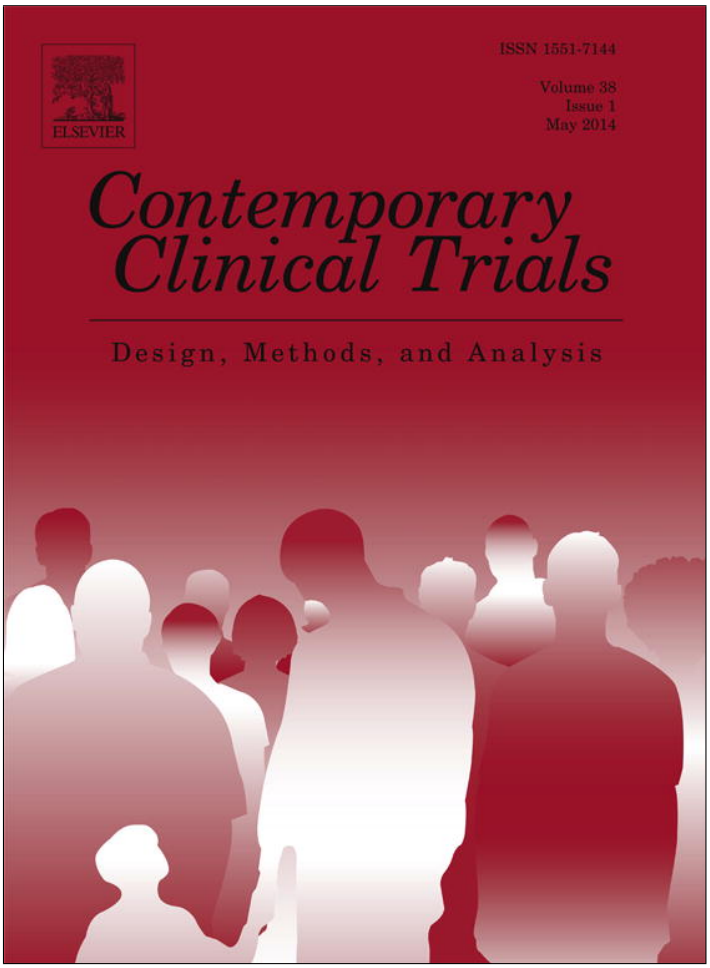
54. Le Reste JY, Nabbe P, Rivet C, Lygidakis C, Doerr C, Czachowski S, et al. The European general practice research network presents the translations of its comprehensive definition of multimorbidity in family medicine in ten European languages. PloS one. 2015; 10(1):e0115796. Epub 2015/01/22. doi: 10.1371/journal.pone.0115796 PMID: 25607642; PubMed Central PMCID: PMCPmc4301631.

55. Vanderver A, Prust M, Tonduti D, Mochel F, Hussey HM, Helman G, et al. Case definition and classification of leukodystrophies and leukoencephalopathies. Molecular genetics and metabolism. 2015; 114 (4):494–500. Epub 2015/02/05. doi: 10.1016/j.ymgme.2015.01.006 PMID: 25649058.

56. San L, Serrano M, Canas F, Romero SL, Sanchez-Cabezudo A, Villar M. Towards a pragmatic and operational definition of relapse in schizophrenia: A Delphi consensus approach. International journal of psychiatry in clinical practice. 2015:1–9. Epub 2014/12/31. doi: 10.3109/13651501.2014.1002501 PMID: 25547440.

57. Post MW. Definitions of quality of life: what has happened and how to move on. Topics in spinal cord injury rehabilitation. 2014; 20(3):167–80. Epub 2014/12/09. doi: 10.1310/sci2003-167 PMID: 25484563; PubMed Central PMCID: PMCPmc4257148.

58. Parnell RW. Health examinations of students; pilot survey in Oxford. Lancet. 1947; 2(6487):939–41. Epub 1947/12/27. PMID: 18897735.



Contemporary Clinical Trials 38 (2014) 130–133

Contents lists available at ScienceDirect

Contemporary Clinical Trials

journal homepage: www.elsevier.com/locate/conclintrial

Pilot and feasibility studies: Is there a difference from each other and from a randomised controlled trial?

Amy L. Whitehead, Benjamin G.O. Sully, Michael J. Campbell *

Design, Trials and Statistics Group, School of Health and Related Research, University of Sheffield, Regent Court, 30 Regent Street, Sheffield S1 4DA, UK

ARTICLE INFO

Article history:
Received 27 January 2014
Received in revised form 3 April 2014
Accepted 5 April 2014
Available online 13 April 2014

Keywords:
Pilot
Feasibility
Terminology
Reporting

ABSTRACT

Background: A crucial part in the development of any intervention is the preliminary work carried out prior to a large-scale definitive trial. However, the definitions of these terms are not clear cut and many authors redefine them. Because of this, the terms *feasibility* and *pilot* are often misused.
Aim: To provide an introduction to the topic area of pilot and feasibility trials and draw together the work of others in the area on defining what is a pilot or feasibility study.
Methods: This study used a review of definitions and advice from the published literature and from funders' websites. Examples are used to show evidence of good practice and poor practice.
Results: We found that researchers use different terms to describe the various stages of the research process. Some define the terms feasibility and pilot as being different whereas others argue that these terms are synonymous. All reflective papers agree that feasibility/pilot studies should not test treatment comparisons nor estimate feasible effect sizes. However, this is not universally observed in practice.
Summary: We believe that the term 'feasibility' should be used as an overarching term for preliminary studies and the term 'pilot' refers to a specific type of study which resembles the intended trial in aspects such as, having a control group and randomisation. However, studies labelled 'pilot' should have different aims and objectives to main trials and also should include an intention for future work. Researchers should not use the title 'pilot' for a trial which evaluates a treatment effect.

1. Introduction

During recent years, there has also been an increasing emphasis on the importance of preliminary work prior to the organisation of large-scale, publicly funded randomised controlled trials. Many large public funding bodies now expect substantial work to have been done prior to the main bid. Some funding streams, such as UK National Institute for Health Research (NIHR) Research for Patient Benefit (RfPB) [1] and the US NIH R34 funding mechanism [2], recognise this through the provision of substantial sums of money to support such work. The value of preliminary work is now recognised and researchers are encouraged to publish their pilot work in advance of their main trial, and some publishers are willing to publish such results. However, there remains much confusion about the purpose of preliminary work and also of terminology used. The NIHR use the terms 'feasibility' and 'pilot' to distinguish between different stages in the research process [3]. Although these terms are frequently used in the literature, they are used inconsistently and interchangeably [4], while other authors choose to use different terms completely to define the stages of development [5].

There is also the temptation to label a trial 'pilot' to excuse a small sample size, or one conducted in one locality, but still with the intention of running a study with treatment comparison as the main objective.

The aim of this paper is to provide an introduction to the topic area of pilot and feasibility trials. We will draw together the work of others that has been done in this area, describing

* Corresponding author. Tel.: +44 114 222 0839.
E-mail address: m.j.campbell@sheffield.ac.uk (M.J. Campbell).

http://dx.doi.org/10.1016/j.cct.2014.04.001
1551-7144/

current definitions, their overlaps and points of divergence. We use examples to illustrate good and poor practice and conclude with some recommendations on the use of the terms. This paper adds to our earlier work [4] by critiquing earlier definitions, and providing examples to support our criticism.

1.1. Current definitions

Within the pharmaceutical sector testing, drug efficacy has long had a tradition of clearly defined stages, from the initial phase 1 first-into-man studies through the phase 4 post-marketing studies. However, for large publicly funded trials, particularly of complex interventions and modes of care, the definitions and stages of trials have been less well defined/clear-cut. There have been several attempts to provide guidance on the definitions of a pilot and feasibility study. A review of papers published in 2001 in seven major journals looked at the objectives of pilot studies in the literature [6] to clarify the definition of pilot study. This was repeated in 2010, and the work extended to distinguish between pilot and feasibility studies in the article search and looking at the components of the studies [4]. The authors of these studies found that studies labelled 'pilot' generally used stricter methodology than studies labelled 'feasibility' and that pilot studies mostly reported their results as inconclusive and suggested further work, whereas feasibility studies did not state the same intention. They argue that the distinction between the two terms is not clear cut. However, they suggest the adoption of the NETSCC (NIHR Evaluation, Trials and Studies Coordinating Centre) definition which does distinguish between the two types of study [3].

The NETSCC [3] define feasibility studies as studies used to estimate important parameters that are needed to design the main study, e.g., standard deviation of the outcome measure, willingness of patients to be randomised, willingness of clinicians to recruit participants, number of people eligible, follow-up rates, response rates and adherence/compliance rates. Feasibility studies may have no plan for further work and their aim is to assess whether it is possible to perform a full-scale study.

The NETSCC [3] define a pilot study as a version of the main study run in miniature to determine whether the components of the main study can all work together. They suggest that a pilot should focus on the processes of running the main study, i.e., to ensure the mechanisms of recruitment, randomisation, treatment and follow-up assessments. The aim of the pilot is to provide training and experience in the running of the trial and to highlight any problems so they may be corrected before the main study begins. There must also be a plan for further work. A pilot study can be either external or internal to the main study.

This latter definition is comparable to the UK NICE definition of a pilot study as 'a small-scale "test" of a particular approach … The aim would be to highlight any problems or areas of concern and amend it before the full-scale study begins [7].'

However, in contrast, Arnold et al. [5] provided three separate definitions for different types of pre-clinical work: pilot work, pilot studies and pilot trials. They defined pilot *work* as 'any background research that informs a future study'; pilot *studies* as 'studies with a specific hypothesis, objective and methodology'; and a pilot *trial* as 'a stand-alone pilot study with a randomisation procedure'. Indeed the authors advocated against using the term *feasibility study*, arguing that it 'does not reflect the scope of many pilot studies'. These definitions differ from most others in that they distinguish between the different possible objectives of pilot studies, but do not include the term feasibility whatsoever. The movement through development stages is defined by using the words; work, study and trial instead of the terms feasibility and pilot.

Thabane et al. [8], in their tutorial on pilot studies, do not distinguish between feasibility and pilot studies and simply note that the terms are used synonymously. They do however note that the main focus of a pilot study should be to test the feasibility of conducting a full study rather than statistical significance, and that many pilot studies fail to recognise this.

Leon et al. [9] state that a pilot study can be used to evaluate the feasibility of recruitment, randomization, retention, assessment procedures, new methods and implementation of the novel intervention. A pilot study is not a hypothesis testing study. Safety, efficacy and effectiveness are not evaluated in a pilot. Contrary to tradition, a pilot study does not provide a meaningful effect size estimate for planning subsequent studies due to the imprecision inherent in data from small samples. Thus, effect sizes provided by pilot studies should not be used to power a subsequent full trial. Instead clinical experience should be used to define a *clinically meaningful* effect. A pilot study is a requisite initial step in exploring a novel intervention or an innovative application of an intervention. Pilot results can inform feasibility and identify modifications needed in the design of a larger, ensuing hypothesis testing study.

This is similar to the British Medical Research Council's (MRC's) complex interventions guidelines, which urge the reader to exercise caution when using the results of a pilot study to make assumptions about the required sample size, likely response rates, etc., when the evaluation is scaled up [10]. These guidelines do not give an exact definition of a pilot or feasibility study; instead, they focus on the outcomes of the feasibility and piloting stage. Investigators should be confident that the intervention can be delivered as intended and be able to make safe assumptions about the effect sizes, variability, recruitment rates and retention to aid in the designing of the main study. They do note that 'a pilot study need not be a "scale model" of the planned main stage evaluation, but should address the main uncertainties that have been identified in the development work'.

1.2. Examples

Krarup et al. [11] describe a trial, the ExSTroke Pilot trial, to examine the benefits of exercise in patients who have had a stroke. They intended to recruit 300 subjects, but this was powered on a postulated difference in treatment groups from a surrogate outcome, the Physical Activity Scale for the Elderly (PACE). The reason for the term 'pilot' in the title could be inferred because the study was not powered for recurrent stroke, MI, or mortality. The results were published [12] as a randomised controlled trial. The trial was criticised because it did not follow guidelines for the developing of complex interventions such as those of the MRC [10], and 'we

might have expected modelling of active ingredients of the intervention (given that it was a pilot study) and testing the feasibility of the approach' [13].

In contrast, the LIFE study [14] is also described as a pilot study. The study intended to recruit 400 adults and the aims were as follows: (a) estimate the sample size needed for a full scale trial, (b) examine the consistency of the effects of the physical activity intervention on several continuous measures of physical function, (c) assess the feasibility of recruitment, (d) evaluate study adherence and retention, (e) evaluate the efficacy of a stepped care approach for managing inter-current illness in this at-risk population and (f) develop a comprehensive system for monitoring and ensuring participant safety. Two points can be made. First, the objectives of the study are consistent with the objectives of a pilot study, except (e) since it was not powered to evaluate efficacy. Second, the size of the projected pilot, at 400, exceeds the size of many full studies and is not justified in relation to the objectives. The outcomes of some of these objectives were subsequently published. For example, the investigators evaluated the longitudinal distributions of four standardised outcomes to contrast how they may serve as primary outcomes of future clinical trials: ability to walk 400 m, ability to walk 4 m in ≤10 s, a physical performance battery and a questionnaire focused on physical function. They concluded that the ability to walk 400 m as a dichotomous outcome provided the smallest sample size projections and that a 4-year trial based on the outcome of the 400-m walk is projected to require $n = 962–2234$ to detect an intervention effect of 30%–20% with 90% power [15]. In fact, they are now running the main study, a trial of 1600 people followed up for 2.7 years [16]. This outcome is entirely coherent with that of the pilot study. However, in view of the size of the pilot, they could not resist also doing some treatment comparisons [17,18]. It is also of note that the size of the pilot was 25% of the main study, which leads one to query the correct ratio in size of the pilot and main study.

Cooper et al. [19] conducted the COSMOS pilot trial to investigate the use of computerised cognitive behavioural therapy (CCBT) for the treatment of depression in patients with multiple sclerosis. They recruited 24 patients based on the precision of estimates to be used to design the main trial [20]. The patients were randomised between CCBT and the usual treatment. The objectives of the study were to estimate the recruitment, withdrawal and dropout rate, sample size estimation and preferences for service delivery (home or elsewhere) and to test the feasibility of recruitment methods, questionnaires and the proposed outcome measures.

This is a good example of a pilot trial. The aims of the trial are consistent with the definition of a pilot trial, outlined in the previous section (to assess the trials processes and procedures, i.e., the questionnaires and the recruitment strategy, and to estimate values for the future trials sample size calculation, i.e., the standard deviation and the dropout rate). Although the effect size was calculated it is not assessed for statistical significance.

The trial was evaluated on the level of recruitment achieved. They had to approach nearly 600 patients to get the 24 in the trial, which resulted in a recruitment rate of 4.1% which was much lower than expected, and only 9/12 (75%) completed at least four out of the eight sessions. The authors concluded that a further trial was not feasible without a change in the eligibility criteria.

2. Discussion

It can be seen that there is still confusion around the terms. Some use the terms feasibility and pilot interchangeably [8] whereas others define the terms separately [3,4,6]. It is problematic to look to the literature to find a difference between pilot and feasibility study as a trial may be labelled as a pilot or feasibility study, but this does not mean that it is a pilot or feasibility study under someone else's definition.

From the review of the literature, we found that the distinguishing features of a pilot study from a feasibility study are as follows:

- Stricter study methodology (e.g., a justification of the sample size)
- An intention for further work
- Smaller version of the main study (e.g., use of a control group and randomisation)
- A focus on trial processes

The stricter methodology may stem from the fact that pilot studies are more likely to mimic the design of the main study, in order to test the processes and provide training to trial staff and alleviate problems before the larger trial. This restriction does not hold for a feasibility study, where a systematic review or meta-analysis may be a feasibility study. A pilot study, apart from investigating how the trial procedures will work in the future trial, may also test the feasibility of a larger study so it could be said that pilot studies are also feasibility studies. However, the inverse cannot be said that all feasibility studies are pilot studies. From this, one could conclude that a pilot study is a special type of feasibility study which has a plan for further work and mimics the envisioned definitive trial. In addition, we could also define a pilot *trial* as a pilot study which also involves randomisation between treatment groups.

The plan for further work is crucial for pilot studies; otherwise, the study may be seen as an underpowered trial which are deemed unethical and have limited scientific use. As we have shown, pilot studies and randomised controlled trials (RCTs) have different aims and objectives [4]. An RCT will test the efficacy of a new intervention, and a pilot study should only test other aspects of the trial design in preparation for this definitive assessment of the treatment. The term 'pilot' implies an intention for further definitive work in the future.

It is impossible to legislate on the use of terminology, but we suggest that if journals and reviewers adopt a more consistent usage, then it would make the reporting and reviewing of such studies much simpler.

It could be argued that trials which use a surrogate end point, such as the ExStroke trial [11] are in fact 'pilot' studies even if they test for treatment comparisons. However, to be consistent with the previous paragraph, they only deserve this label if there are clear criteria to decide on a whether to conduct a subsequent trial using clinically meaningful outcomes, and a clear intention of conducting such a trial if the criteria are met. Otherwise, the title should clearly define the trial as one that uses surrogate end points. Thus, the ExStroke trial could have specified what size difference in the PACE outcome would have justified further follow up for stroke and death, or an extension of the trial to include these outcomes.

Thought should also go into whether a pilot study should be 'internal' or 'external'. Parameters such as the recruitment rate could be determined within the main trial, with a time point specified as to when a decision would be made to continue or abandon the trial, depending on the number of patients recruited. There is a balance to be had. In external pilot studies, patients are 'wasted' in that they do not contribute to the final clinical end point, which may be a problem if patients are hard to come by. On the other hand, in an internal pilot, patients are also 'wasted' if the trial is abandoned. A further point is that it may be difficult to get funding to conduct a pilot trial whose outcome is something like recruitment rates and appears unrelated to patient benefit. However, funding an external pilot, and with contingent funding for the main trial may be easier for the funding agency than committing funding for the whole trial, with contingent withdrawal of funding if progress is unsatisfactory.

3. Conclusion

The distinction between pilot and feasibility studies is still a very grey area, with various definitions having been suggested by clinical trial methodology researchers. We suggest it is futile to ascribe a particular meaning to the term 'feasibility' and that all preliminary trial work could be described as 'feasibility'; therefore, it could be thought of as a catch-all term for preliminary work. However, the term 'pilot' could be reserved for a study that mimics the definitive trial design in that it may include control groups and randomisation but whose explicit objective is *not* to compare treatment groups but rather to ensure the main trial delivers maximum benefit. Trials that use surrogate end points should be described as pilot trials only if they include clear criteria for proceeding to a main trial.

Acknowledgements

We thank members of the CONSORT team on Pilot and Feasibility trials for helpful discussion (Sandra Eldridge, Gill Lancaster, Christine Bond, Sally Hopewell and Lehana Thabane).

References

[1] NIHR Research for Patient Benefit Programme. Director's Message 6 - Thinking about applying for funding for a pilot study? [accessed December 18 2013] http://www.ccf.nihr.ac.uk/RfPB/Documents/RfPB_Directors_message_6.pdf.
[2] National Institute of Mental Health. Pilot Intervention and Services Research Grants (R34). [accessed December 18 2013] http://grants.nih.gov/grants/guide/pa-files/PAR-09-173.html.
[3] NETSCC. Glossary: Feasibility and Pilot Studies. [accessed December 18 2013] http://www.netscc.ahttp://www.nets.nihr.ac.uk/glossary?result_1655_result_page=Pc.uk/glossary/; 2013.
[4] Arain M, Campbell MJ, Cooper CL, Lancaster GA. What is a pilot or feasibility study? A review of current practice and editorial policy. BMC Med Res Methodol 2010;10:67. http://dx.doi.org/10.1186/1471-2288-10-67.
[5] Arnold DM, Burns KE, Adhikari NK, Kho ME, Meade MO, Cooke DJ. The design and interpretation of pilot trials in clinical research in critical care. Crit Care Med 2009;37(1):S69–74.
[6] Lancaster GA, Dodd S, Williamson PR. Design and analysis of pilot studies: recommendations for good practice. J Eval Clin Pract 2004;10(2):307–12.
[7] NICE. Glossary. [accessed December 19 2013] http://www.nice.org.uk/website/glossary/glossary.jsp?alpha=P.
[8] Thabane L, Ma J, Chu R, Cheng J, Ismaila A, Rios LP, et al. A tutorial on pilot studies: the what, why and how. BMC Med Res Methodol 2010;10. http://dx.doi.org/10.1186/1471-2288-10-1.
[9] Leon AC, Davis LL, Kraemer HC. The role and interpretation of pilot studies in clinical research. J Psychiatr Res 2011;45(5):626–9.
[10] Medical Research Council. Developing and Evaluating Complex Interventions: New guidance. [accessed December 18 2013] http://www.mrc.ac.uk/Utilities/Documentrecord/index.htm?d=MRC004871.
[11] Krarup L-H, Gluud C, Truelsen T, Pedersen A, Lindahl M, Hansen L, et al. The ExSTroke Pilot Trial: rationale, design, and baseline data of a randomized multicenter trial comparing physical training versus usual care after an ischemic stroke. Contemp Clin Trials 2008;29(3):410–7.
[12] Boysen G, Krarup L-H, Zeng X, Oskedra A, Kõrv J, Andersen G, et al. ExStroke Pilot Trial of the effect of repeated instructions to improve physical activity after ischaemic stroke: a multinational randomised controlled clinical trial. BMJ 2009;339:b2810. http://dx.doi.org/10.1136/bmj.b2810.
[13] Mutrie N. ExStroke Pilot Trial of the effect of repeated instructions to improve physical activity after ischaemic stroke: a multinational randomised controlled clinical trial: Response. http://www.bmj.com/content/339/bmj.b2810?tab=responses . [accessed 19th Dec 2013].
[14] Rejeski WJ, Fielding RA, Blair SB, Guralnik JM, Gill TM, Hadley EC, et al. The lifestyle interventions and independence for elders (LIFE) pilot study: design and methods. Contemp Clin Trials 2005;26:141–54.
[15] Espeland MA, Gill TM, Guralnik J, Miller ME, Fielding R, Newman AB, et al. Designing clinical trials of interventions for mobility disability: results from the lifestyle interventions and independence for elders pilot (LIFE-P) trial. J Gerontol A Biol Sci Med Sci 2007;62(11):1237–43.
[16] The Life Study. General Public Information. [accessed 19th Dec 2013] https://www.thelifestudy.org/public/index.cfm; 2013.
[17] Pahor M, Blair SN, Espeland M, Fielding R, Gill TM, Guralnik JM, et al. Effects of a physical activity intervention on measures of physical performance: results of the lifestyle interventions and independence for Elders Pilot (LIFE-P) study. J Gerontol A Biol Sci Med Sci 2006;61(11):1157–65.
[18] Rejeski WJ, Marsh AP, Chmelo E, Prescott AJ, Dobrosielski M, Walkup MP, et al. The lifestyle interventions and independence for elders pilot (LIFE-P): 2-year follow-up. J Gerontol A Biol Sci Med Sci 2009;64(4):462–7.
[19] Cooper CL, Hind D, Dimairo M, Thake A, Parry GD, O'Cathain A, et al. Computerised cognitive behavioural therapy for the treatment of depression in people with multiple sclerosis: external pilot trial. Trials 2011;12:259. http://dx.doi.org/10.1186/1745-6215-12-259.
[20] Julious SA. Sample size of 12 per group rule of thumb for a pilot study. Pharm Stat 2005;4:287–91.

CONSORT 2010 statement: extension to randomised pilot and feasibility trials

Sandra M Eldridge,[1] Claire L Chan,[1] Michael J Campbell,[2] Christine M Bond,[3] Sally Hopewell,[4] Lehana Thabane,[5] Gillian A Lancaster[6] on behalf of the PAFS consensus group

[1]Centre for Primary Care and Public Health, Queen Mary University of London, London, UK
[2]School of Health and Related Research, University of Sheffield, Sheffield, UK
[3]Centre of Academic Primary Care, University of Aberdeen, Aberdeen, Scotland, UK
[4]Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK
[5]Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada
[6]Department of Mathematics and Statistics, Lancaster University, Lancaster, UK

Correspondence to: S M Eldridge s.eldridge@qmul.ac.uk

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2016;355:i5239
http://dx.doi.org/10.1136/bmj.i5239

Accepted: 18 September 2016

The Consolidated Standards of Reporting Trials (CONSORT) statement is a guideline designed to improve the transparency and quality of the reporting of randomised controlled trials (RCTs). In this article we present an extension to that statement for randomised pilot and feasibility trials conducted in advance of a future definitive RCT. The checklist applies to any randomised study in which a future definitive RCT, or part of it, is conducted on a smaller scale, regardless of its design (eg, cluster, factorial, crossover) or the terms used by authors to describe the study (eg, pilot, feasibility, trial, study). The extension does not directly apply to internal pilot studies built into the design of a main trial, non-randomised pilot and feasibility studies, or phase II studies, but these studies all have some similarities to randomised pilot and feasibility studies and so many of the principles might also apply.

The development of the extension was motivated by the growing number of studies described as feasibility or pilot studies and by research that has identified weaknesses in their reporting and conduct. We followed recommended good practice to develop the extension, including carrying out a Delphi survey, holding a consensus meeting and research team meetings, and piloting the checklist.

The aims and objectives of pilot and feasibility randomised studies differ from those of other randomised trials. Consequently, although much of the information to be reported in these trials is similar to those in randomised controlled trials (RCTs) assessing effectiveness and efficacy, there are some key differences in the type of information and in the appropriate interpretation of standard CONSORT reporting items. We have retained some of the original CONSORT statement items, but most have been adapted, some removed, and new items added. The new items cover how participants were identified and consent obtained; if applicable, the prespecified criteria used to judge whether or how to proceed with a future definitive RCT; if relevant, other important unintended consequences; implications for progression from pilot to future definitive RCT, including any proposed amendments; and ethical approval or approval by a research review committee confirmed with a reference number.

This article includes the 26 item checklist, a separate checklist for the abstract, a template for a CONSORT flowchart for these studies, and an explanation of the changes made and supporting examples. We believe that routine use of this proposed extension to the CONSORT statement will result in improvements in the reporting of pilot trials.

Editor's note: In order to encourage its wide dissemination this article is freely accessible on the *BMJ* and *Pilot and Feasibility Studies* journal websites.

The Consolidated Standards of Reporting Trials (CONSORT) statement (www.consort-statement.org) is a guideline designed to improve the transparency and quality of the reporting of randomised trials. It was first published in 1996, revised in 2001, last updated in 2010,[1 2] and published simultaneously in 10 leading medical journals, including the *Lancet*, *JAMA*, *BMJ*, *Annals of Internal Medicine*, and *PLoS Medicine*. The CONSORT statement comprises a checklist of the minimum essential items that should be included in reports of randomised trials and a diagram documenting the flow of participants through the trial.

The development of CONSORT guidelines has received considerable international recognition. The CONSORT statement has been cited more than 8000 times and has received support from the World Association of Medical Editors, Council of Science Editors, International Committee of Medical Journal Editors, and more than 600 journals worldwide. Several studies have examined the impact of the statement on the reporting quality of published randomised trials and found that adoption of the statement leads to an increase in reporting quality.[3]

In addition to the CONSORT statement, extensions to the CONSORT checklist for reporting trials with non-inferiority, equivalence, and cluster or pragmatic designs have been published,[4–6] as have extension checklists for reporting harms,[7] different types of interventions (non-drug treatments[8] and herbal interventions[9]), and patient reported outcomes.[10] The main CONSORT statement and all of the current extensions focus on trials for which the research question centres on the effectiveness or efficacy of an intervention. However, some randomised trials, that we refer to as pilot and feasibility trials, do not have effectiveness or efficacy as their primary focus. Rather, they are designed to support the development of a future definitive RCT. By "definitive" in this context we mean an appropriately powered study focusing on effectiveness or efficacy. The need for high standards in conduct and reporting applies just as much to pilot and feasibility trials as it does to definitive trials.

Scope of this paper

In this article we present an extension to the CONSORT statement for randomised pilot and feasibility trials conducted in advance of a future definitive RCT. In keeping with the broad scope of CONSORT, the future definitive RCT might evaluate either the efficacy or the effectiveness of an intervention. The primary aim of the randomised pilot or feasibility trial, however, is to assess feasibility of conducting the future definitive RCT.

We make no distinction in this extension between pilot and feasibility randomised trials. Although in practice we recognise that different researchers might have preferences for different terms, the lack of distinction is based on a framework developed by the authors, which defines such studies.[11] In that framework, a feasibility study for a future definitive RCT asks whether the future trial can be done, should be done, and, if so, how. Pilot studies are a subset of feasibility studies. They ask the same questions about feasibility (whether the future trial can be done, should be done, and, if so, how) but have a particular design feature: in a pilot study (that might or might not be randomised) the future definitive RCT, or part of it, is conducted on a smaller scale.

For brevity, we use the term "pilot trial" to refer to any randomised study in which a future definitive RCT, or a part of it, is conducted on a smaller scale. However, these studies might legitimately be referred to using any of the following terms: pilot RCT, randomised pilot trial, pilot trial, pilot study, randomised pilot study, feasibility RCT, randomised feasibility trial, feasibility trial, feasibility study, or randomised feasibility study. In fact, we have set no restrictions on the terminology used to describe pilot trials; rather we have specified only that they are randomised, conducted in advance of a future definitive RCT, and primarily aim to assess feasibility.

The development of this extension was motivated by the growing number of studies described as feasibility or pilot studies[12] and by research that has identified weaknesses in the reporting and conduct of these studies.[12–15] We expect that improved reporting quality will lead to more high quality examples of pilot trials, enabling yet further improvements in the conduct of pilot trials and making it possible for readers to use the results of reported pilot studies in preparing future trials in similar settings and with similar participants. Because the purpose of a pilot trial (to assess feasibility) is different from that of the future definitive RCT (to assess effectiveness or efficacy), the focus of the reporting should be different, and that difference is reflected in the extension.

The extension does not apply to internal pilot studies that are built into the design of a main trial, or to non-randomised pilot and feasibility studies. However, much of what is presented here might apply to, or be adapted to apply to, these types of pilot or feasibility studies or similar types of trial, such as "proof of concept" or phase II trials done in the development of drugs.[16 17] Proof of concept or phase II trials are small RCTs the main objective of which are to inform the sponsor whether or not to continue the development of a drug with larger trials. Similar to pilot trials, the focus is on assessing the feasibility of further development rather than assessing effectiveness or efficacy. However, to do this these trials tend to focus on aspects such as safety and potential effectiveness or efficacy. They might use accepted methods devised for phase II trials[18] to assess the outcome to be used in a future phase III trial (which could be meta-analysed if required)[19] or use surrogate outcomes—that is, intermediate measures, often biochemical, which have less direct impact on a patient than, for example, cure or death, but which should be associated with these "hard" outcomes. Safety, and potential effectiveness or efficacy, are usually less important in pilot trials, where the focus is on the development of interventions and their evaluation and where issues related to feasibility might be different. Nevertheless, pilot trials do sometimes assess

doi: 10.1136/bmj.i5239 | *BMJ* 2016;355:i5239 | the bmj

potential effectiveness using surrogate outcomes. For example, oxygenation of the blood as a surrogate measure for improved lung function and survival[20] or the number of steps walked each day as a surrogate for clinical measures of heart disease.[21]

Here we present an extension to the standard CONSORT guidelines for reporting RCTs. Many investigators, however, use qualitative research alongside other methods to assess feasibility. The amount of qualitative work conducted at the pilot and feasibility stage, its relation with any pilot trial, and the way investigators want to report this work, varies. Stand-alone qualitative studies that are reported separately from the pilot trial, such as Hoddinott et al and Schoultz et al,[22 23] should follow appropriate reporting guidelines[24–26] and should provide link references to other pilot work carried out in preparation for the same definitive trial. When qualitative work is reported within the primary report of a pilot trial,[27] it is not always possible to put sufficient detail into the methods section of the report to comply with reporting guidelines for qualitative studies. If this is the case, we recommend an online supplement or appendix to report the methods in detail. O'Cathain et al, Hoddinott et al, and Schoultz et al have provided guidelines and examples for conducting qualitative feasibility studies alongside pilot trials.[22 23 26 28 29]

Adapting the CONSORT statement for pilot trials

The development of this CONSORT extension for pilot trials is described briefly here and in detail elsewhere.[30] Before developing the checklist for this extension, the research team agreed on the definitions of pilot and feasibility studies. This was done by initially considering pilot and feasibility studies to be discrete types of study and therefore in need of separate checklists. However, preliminary work concluded that pilot and feasibility studies could not be defined in a mutually exclusive way, compatible with current understanding and the use of these terms among the research community. We therefore adopted an overarching definition of feasibility studies, with pilot studies being a subset, and developed a single checklist for such studies that use a randomised approach, referred to as pilot trials in this paper. The process of agreeing on the definitions of feasibility and pilot studies and the underpinning conceptual framework are reported separately.[11] That work was done in parallel with the development of the checklist (table 1). We used the principles in box 1 to guide the work.

In stage 1, the research team met and worked through each of the existing CONSORT checklist items, agreeing whether each was relevant and should be retained, not relevant and should be excluded, or needed rewording in the context of either a feasibility study or pilot study. This resulted in two checklists. We then applied the revised checklists to a sample of 30 articles identified from previous work[13 15] and our own personal collections.

In stage 2, we used a modified Delphi survey to seek consensus on the appropriateness of each of the checklist items. Participants (n=93) were asked to rate each item on a scale of 1 to 9 (1=not at all appropriate to 9=completely appropriate). They were also given the opportunity to comment on each item, definitions of pilot and feasibility studies, and the perceived usefulness of the checklist.[11]

In stage 3, participants in the Delphi survey were asked to review responses for items that 70% or more of participants had rated as 8 or 9 in round 1 of the survey and to make additional comments on these items. They were asked to review the remaining items and classify each using one of four options: discard, keep, unsure, or no opinion. They were also asked to add any items they believed had been missed. In total 93/120 (77.5%) responses were received for round 1 and 79/93 (84.9%) for round 2.

In stage 4, the research team met face to face to review the feedback from the Delphi survey and to revise the checklist. In stage 5, the revised checklist was then further reviewed in detail during a two day expert

Table 1 | Stages of adapting CONSORT statement for pilot trials

Stage	Activity	Participants	Venue (or virtual meeting)	Date
1	Drafting of definitions and preliminary adaptation of CONSORT checklist items	Research team	London	Dec 2012
2	1st round of modified Delphi process using online administration	Invited experts from research community (trialists, methodologists, statisticians, funders, and journal editors)	Email distribution	Jul-Aug 2013
3	2nd round of modified Delphi process	As for round 1	Email distribution	Sept-Oct 2013
4	Review of results from Delphi process and redrafting checklist	Research team	London	Feb 2014
5	Consensus meeting	Invited experts (trialists, methodologists, statisticians, funders, journal editors, and members of CONSORT executive)	Oxford	Oct 2014
6	Review of consensus meeting feedback and drafting final checklist	Research team	Email consultation with consensus participants; and meetings in London	Dec 2014-Dec 2015; and Jan, Jun, Dec 2015
7	Further review and piloting	Research team	Email consultation with consensus participants; and piloting by independent researchers writing up pilot studies	Mar 2016; and Jan-Mar 2016

Box 1: Methodological considerations and principles that guided the development of the CONSORT extension to pilot trials

- The rationale of a pilot trial is to investigate areas of uncertainty about the future definitive RCT
- The primary aims and objectives of a pilot trial are therefore about feasibility, and this should guide the methodology used in the pilot trial
- Assessments or measurements to address each pilot trial objective should be the focus of data collection and analysis. This might include outcome measures likely to be used in the definitive trial but, equally, it might not
- Since the aim of a pilot trial is to assess the feasibility of proceeding to the future definitive RCT, a decision process about how to proceed needs to be built into the design of the pilot trial. This might involve formal progression criteria to decide whether to proceed, to proceed with amendments, or not to proceed
- Methods used to address each pilot trial objective can be qualitative or quantitative. A mixed methods approach could result in both types of data being reported within the same paper. Equally, a process evaluation or other qualitative study can be done alongside a pilot trial and reported separately in more detail
- The number of participants in a pilot study should be based on the feasibility objectives and some rationale should be given
- Formal hypothesis testing for effectiveness (or efficacy) is not recommended. The aim of a pilot trial is not to assess effectiveness (or efficacy) and it will usually be underpowered to do this

consensus meeting. In stage 6, some checklist items were reworded to ensure clarity of meaning and purpose, and the research team met face to face a further three times to agree on the final wording of the checklist, identify examples of good reporting, and develop the explanation and elaboration section of this paper. A full draft of the paper was then sent to members of the consensus meeting to ensure it fully reflected the discussion of the meeting.

Table 2 presents the final checklist, laid out in accordance with other CONSORT extensions. Items in the standard checklist column should be adhered to unless the extension column indicates a change in the item. Box 1 lists the methodological considerations and principles that guided the process.

Extension of CONSORT 2010 to pilot trials

Title and abstract

- Item 1a

- *Standard CONSORT item*: identification as a randomised trial in the title

- *Extension for pilot trials*: identification as a pilot or feasibility randomised trial in the title

- *Example 1 (using the words pilot, randomised, and trial)*

- "Bespoke smoking cessation for people with severe mental ill health (SCIMITAR): a pilot randomised controlled trial"[31]

- *Example 2 (using the words feasibility, randomised, and trial)*

- "A cluster randomised feasibility trial evaluating nutritional interventions in the treatment of malnutrition in care home adult residents"[32]

- *Explanation*

The primary focus of these guidelines is randomised pilot and feasibility trials. To ensure that these types of studies can be easily identified from specific search criteria, a title containing the descriptors "pilot" or "feasibility" as well as "randomised" provides a necessary, recognised terminology for selecting randomised pilot and feasibility trials.[13] This would also enable these studies to be easily indexed in electronic databases, such as PubMed.[33] Although the descriptors might appear in the title for many studies, they might not necessarily occur together, as in: "Feasibility of a randomised trial of a continuing medical education program in shared decision-making on the use of antibiotics for acute respiratory infections in primary care: the DECISION+ pilot trial."[34] Furthermore, in some cases authors might use the phrase "randomised pilot study" or "randomised feasibility study," as in "'Not just another walking program': Everyday Activity Supports You (EASY) model—a randomized pilot study for a parallel randomized controlled trial."[21] Such papers could be identified in appropriate searches. However, in general we recommend the descriptors are given together in one phrase, and the word "trial" rather than "study" is used, as in "randomised pilot trial" or "randomised feasibility trial."

- Item 1b

- *Standard CONSORT item*: structured summary of trial design, methods, results, and conclusions (for specific guidance see CONSORT for abstracts)[35 36]

- *Extension for pilot trials*: structured summary of pilot trial design, methods, results, and conclusions (for specific guidance see CONSORT abstract extension for pilot trials) (table 3)

- *Example*

See figures 1 to 3.[21]

- *Explanation*

Abstracts can follow different structures dependent on a journal's style. They are typically around 300 words. We outline what information should be reported in the abstract irrespective of style. This information may also be used for writing conference abstracts. The structure of the abstract does not differ in format from item 1b of the standard CONSORT 2010 guidelines. However, its content focuses on the aims and objectives of the pilot trial and not on the future definitive RCT.

It is important that the abstract contains pertinent information on the background, methods, results, and conclusions in relation to the feasibility objectives and outcomes, and that it states the study is a "randomised" pilot trial. This will aid researchers in understanding the nature of the paper and facilitates electronic searching through the inclusion of specific key words. A statement in the abstract that this study is in preparation for a future definitive RCT is recommended to place it in context. A description of the areas of uncertainty to be addressed and a statement of the feasibility aims and objectives should be included in the background, how these objectives have been addressed in the methods, and results for each objective in the results. If there are a limited number of pilot trial objectives then all should be listed and results for each reported. If there are many pilot trial objectives, then agreement should be reached a priori about which are the most important, to decide whether to proceed to a future definitive RCT, and only these objectives should be reported. An explicit state-

doi: 10.1136/bmj.i5239 | *BMJ* 2016;355:i5239 | the bmj

Table 2 | CONSORT checklist of information to include when reporting a pilot trial

Section/topic and item No	Standard checklist item	Extension for pilot trials	Page No where item is reported
Title and abstract			
1a	Identification as a randomised trial in the title	Identification as a pilot or feasibility randomised trial in the title	
1b	Structured summary of trial design, methods, results, and conclusions (for specific guidance see CONSORT for abstracts)	Structured summary of pilot trial design, methods, results, and conclusions (for specific guidance see CONSORT abstract extension for pilot trials)	
Introduction			
Background and objectives:			
2a	Scientific background and explanation of rationale	Scientific background and explanation of rationale for future definitive trial, and reasons for randomised pilot trial	
2b	Specific objectives or hypotheses	Specific objectives or research questions for pilot trial	
Methods			
Trial design:			
3a	Description of trial design (such as parallel, factorial) including allocation ratio	Description of pilot trial design (such as parallel, factorial) including allocation ratio	
3b	Important changes to methods after trial commencement (such as eligibility criteria), with reasons	Important changes to methods after pilot trial commencement (such as eligibility criteria), with reasons	
Participants:			
4a	Eligibility criteria for participants		
4b	Settings and locations where the data were collected		
4c		How participants were identified and consented	
Interventions:			
5	The interventions for each group with sufficient details to allow replication, including how and when they were actually administered		
Outcomes:			
6a	Completely defined prespecified primary and secondary outcome measures, including how and when they were assessed	Completely defined prespecified assessments or measurements to address each pilot trial objective specified in 2b, including how and when they were assessed	
6b	Any changes to trial outcomes after the trial commenced, with reasons	Any changes to pilot trial assessments or measurements after the pilot trial commenced, with reasons	
6c		If applicable, prespecified criteria used to judge whether, or how, to proceed with future definitive trial	
Sample size:			
7a	How sample size was determined	Rationale for numbers in the pilot trial	
7b	When applicable, explanation of any interim analyses and stopping guidelines		
Randomisation:			
Sequence generation:			
8a	Method used to generate the random allocation sequence		
8b	Type of randomisation; details of any restriction (such as blocking and block size)	Type of randomisation(s); details of any restriction (such as blocking and block size)	
Allocation concealment mechanism:			
9	Mechanism used to implement the random allocation sequence (such as sequentially numbered containers), describing any steps taken to conceal the sequence until interventions were assigned		
Implementation:			
10	Who generated the random allocation sequence, enrolled participants, and assigned participants to interventions		
Blinding:			
11a	If done, who was blinded after assignment to interventions (eg, participants, care providers, those assessing outcomes) and how		
11b	If relevant, description of the similarity of interventions		
Analytical methods:			
12a	Statistical methods used to compare groups for primary and secondary outcomes	Methods used to address each pilot trial objective whether qualitative or quantitative	
12b	Methods for additional analyses, such as subgroup analyses and adjusted analyses	Not applicable	

Table 2 | CONSORT checklist of information to include when reporting a pilot trial

Section/topic and item No	Standard checklist item	Extension for pilot trials	Page No where item is reported
Results			
Participant flow (a diagram is strongly recommended):			
13a	For each group, the numbers of participants who were randomly assigned, received intended treatment, and were analysed for the primary outcome	For each group, the numbers of participants who were approached and/or assessed for eligibility, randomly assigned, received intended treatment, and were assessed for each objective	
13b	For each group, losses and exclusions after randomisation, together with reasons		
Recruitment:			
14a	Dates defining the periods of recruitment and follow-up		
14b	Why the trial ended or was stopped	Why the pilot trial ended or was stopped	
Baseline data:			
15	A table showing baseline demographic and clinical characteristics for each group		
Numbers analysed:			
16	For each group, number of participants (denominator) included in each analysis and whether the analysis was by original assigned groups	For each objective, number of participants (denominator) included in each analysis. If relevant, these numbers should be by randomised group	
Outcomes and estimation:			
17a	For each primary and secondary outcome, results for each group, and the estimated effect size and its precision (such as 95% confidence interval)	For each objective, results including expressions of uncertainty (such as 95% confidence interval) for any estimates. If relevant, these results should be by randomised group	
17b	For binary outcomes, presentation of both absolute and relative effect sizes is recommended	Not applicable	
Ancillary analyses:			
18	Results of any other analyses performed, including subgroup analyses and adjusted analyses, distinguishing prespecified from exploratory	Results of any other analyses performed that could be used to inform the future definitive trial	
Harms:			
19	All important harms or unintended effects in each group (for specific guidance see CONSORT for harms)		
19a		If relevant, other important unintended consequences	
Discussion			
Limitations:			
20	Trial limitations, addressing sources of potential bias, imprecision, and, if relevant, multiplicity of analyses	Pilot trial limitations, addressing sources of potential bias and remaining uncertainty about feasibility	
Generalisability:			
21	Generalisability (external validity, applicability) of the trial findings	Generalisability (applicability) of pilot trial methods and findings to future definitive trial and other studies	
Interpretation:			
22	Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence	Interpretation consistent with pilot trial objectives and findings, balancing potential benefits and harms, and considering other relevant evidence	
22a		Implications for progression from pilot to future definitive trial, including any proposed amendments	
Other information			
Registration:			
23	Registration number and name of trial registry	Registration number for pilot trial and name of trial registry	
Protocol:			
24	Where the full trial protocol can be accessed, if available	Where the pilot trial protocol can be accessed, if available	
Funding:			
25	Sources of funding and other support (such as supply of drugs), role of funders		
26		Ethical approval or approval by research review committee, confirmed with reference number	

doi: 10.1136/bmj.i5239 | *BMJ* 2016;355:i5239 | the bmj

Table 3 | Extension of CONSORT for abstracts for reporting pilot trials

Item	Standard checklist item	Extension for pilot trials
Title	Identification of study as randomised	Identification of study as randomised pilot or feasibility trial
Trial design	Description of the trial design (eg, parallel, cluster, non-inferiority)	Description of pilot trial design (eg, parallel, cluster)
Methods:		
Participants	Eligibility criteria for participants and the settings where the data were collected	Eligibility criteria for participants and the settings where the pilot trial was conducted
Interventions	Interventions intended for each group	
Objective	Specific objective or hypothesis	Specific objectives of the pilot trial
Outcome	Clearly defined primary outcome for this report	Prespecified assessment or measurement to address the pilot trial objectives*
Randomisation	How participants were allocated to interventions	
Blinding (masking)	Whether or not participants, caregivers, and those assessing the outcomes were blinded to group assignment	
Results:		
Numbers randomised	Number of participants randomised to each group	Number of participants screened and randomised to each group for the pilot trial objectives*
Recruitment	Trial status†	
Numbers analysed	Number of participants analysed in each group	Number of participants analysed in each group for the pilot objectives*
Outcome	For the primary outcome, a result for each group and the estimated effect size and its precision	Results for the pilot objectives, including any expressions of uncertainty*
Harms	Important adverse events or side effects	
Conclusions	General interpretation of the results	General interpretation of the results of pilot trial and their implications for the future definitive trial
Trial registration	Registration number and name of trial register	Registration number for pilot trial and name of trial register
Funding	Source of funding	Source of funding for pilot trial

*Space permitting, list all pilot trial objectives and give the results for each. Otherwise, report those that are a priori agreed as the most important to the decision to proceed with the future definitive RCT.
†For conference abstracts.

ment relating to whether the future definitive RCT is likely to go ahead on the basis of the results of the pilot trial should also form part of the discussion and conclusions.

Introduction

- Item 2a

- *Standard CONSORT item*: scientific background and explanation of rationale

- *Extension for pilot trials*: scientific background and explanation of rationale for future definitive trial, and reasons for randomised pilot trial

- *Example*

"Reduced fetal movements (RFM) is a frequently seen problem in maternity care with 6-15% of women reporting attending at least one occasion of RFM to health professionals in the third trimester of pregnancy. RFM, defined by maternal perception of significantly reduced or absent fetal activity, is associated with increased risk of stillbirth and fetal growth restriction (FGR) due to placental dysfunction. Despite this association there is a paucity of evidence to direct clinical management of women presenting with RFM. This has been recently highlighted by guidelines from the Royal College of Obstetrics and Gynaecology (RCOG) and a meta-analysis . . . The absence of high-quality evidence has led to wide variation in management strategies for RFM in high-income settings . . . Although there are randomised controlled trials (RCT) of counting fetal movements by a formal structure (e.g. count to ten) there have been no published RCTs of patient management following presentation with RFM. To undertake an RCT of patient management raises important practical concerns including: maternal anxiety for fetal wellbeing, the need to make a decision regarding participation in a short period of time due to the acute nature of RFM and adherence to protocol. Thus, studies have adopted an approach of changing practice at the unit level in quality-improvement projects or stepwise cluster RCT (AFFIRM, NCT01777022). We performed this study to address whether an RCT of the management of RFM in individual patients was an appropriate trial design, and was feasible with regard to i) maternal recruitment and retention ii) patient acceptability, iii) adherence to protocol. In addition, we wished to confirm the prevalence of poor perinatal outcomes in the study population."[37]

- *Explanation*

It is important that the scientific background sets the scene and gives the rationale and justification for the future definitive RCT and why the pilot trial is needed, because under the principles of the Helsinki declaration it is unethical to expose people unnecessarily to the risks of research.[38] The background and rationale are nicely illustrated in the example. Other related publications, or preliminary work such as systematic reviews, qualitative studies, or additional feasibility work, or absence of such work because no one has looked at this topic before, should also be mentioned. The rationale for the randomised pilot trial should be clearly outlined, including the areas of uncertainty that need to be addressed before the future definitive RCT can take place and why such a trial is needed before proceeding to the future definitive RCT. This rationale is usually reported in the final paragraph of the introduction or background section to provide a justification for the pilot trial.

Original abstract – checklist items reported by the authors are in green

Title: "Not just another walking program": Everyday Activity Supports You (EASY) model – a randomized pilot study for a parallel randomized controlled trial

Background: Maintaining physical activity is an important goal with positive health benefits, yet many people spend most of their day sitting. Our Everyday Activity Supports You (EASY) model aims to encourage movement through daily activities and utilitarian walking. The primary objective of this phase was to test study feasibility (recruitment and retention rates) for the EASY model.
Methods: This 6-month study took place in Vancouver, Canada, from May to December 2013, with data analyses in February 2014. Participants were healthy, inactive, community-dwelling women aged 55–70 years. We recruited through advertisements in local community newspapers and randomized participants using a remote web service. The model included the following: group-based education and social support, individualized physical activity prescription (called Activity 4-1-1), and use of a Fitbit activity monitor. The control group received health-related information only. The main outcome measures were descriptions of study feasibility (recruitment and retention rates). We also collected information on activity patterns (ActiGraph GT3X+ accelerometers) and health-related outcomes such as body composition (height and weight using standard techniques), blood pressure (automatic blood pressure monitor), and psychosocial variables (questionnaires).
Results: We advertised in local community newspapers to recruit participants. Over 3 weeks, 82 participants telephoned; following screening, 68% (56/82) met the inclusion criteria and 45% (25/56) were randomized by remote web-based allocation. This included 13 participants in the intervention group and 12 participants in the control group (education). At 6 months, 12/13 (92%) intervention and 8/12 (67%) control participants completed the final assessment. Controlling for baseline values, the intervention group had an average of 2,080 [95% confidence intervals (CIs) 704, 4,918] more steps/day at 6 months compared with the control group. There was an average between group difference in weight loss of −4.3 [95% CI −6.22, −2.40] kg and reduction in diastolic blood pressure of −8.54 [95% CI −16.89, −0.198] mmHg, in favor of EASY.
Conclusions: The EASY pilot study was feasible to deliver; there was an increase in physical activity and reduction in weight and blood pressure for intervention participants at 6 months.
Trial registration: ClinicalTrials.gov identifier: NCT01842061

Pilot and Feasibility Studies. 2015, 1:4. doi:10.1186/2055-5784-1-4
Word count: 335

Item	Extension for pilot trials	Reported
Title	Identification of study as randomised pilot trial	✓
Trial design	Description of pilot trial design (e.g. parallel, cluster)	x
METHODS		
Participants	Eligibility criteria for participants and the settings where the pilot trial was conducted	✓
Interventions	*Interventions intended for each group*	✓
Objective	Specific objectives of the pilot trial	✓
Outcome	Pre-specified assessment or measurement to address the pilot trial objective(s)[1]	✓
Randomisation	*How participants were allocated to interventions*	✓
Blinding (masking)	*Whether or not participants, care givers, and those assessing the objectives were blinded to group assignment*	x
RESULTS		
Numbers randomised	Number of participants screened and randomised to each group for the pilot trial objective(s)[1]	✓
Recruitment	*Trial status*[2]	*N/A*
Numbers analysed	Number of participants analysed in each group for the pilot objective(s)[1]	✓
Outcome	Results for the pilot objective(s); including any expressions of uncertainty[1]	partly
Harms	*Important adverse events or side-effects*	x
Conclusions	General interpretation of the results of pilot trial and their implications for the future definitive trial	partly
Trial registration	Registration number for pilot trial and name of trial register	✓
Funding	Source of funding for pilot trial	x

Items above in italics are pilot trial checklist items unchanged from the standard RCT checklist.
[1] Space permitting, list all pilot trial objectives and give the results for each. Otherwise, report those which are a priori agreed as the most important (main) to the decision to proceed with the future definitive trial. [2] For conference abstracts.

Fig 1 | Example of abstract for report of pilot trial,[21] shown alongside CONSORT for abstracts extension for pilot randomised trials

Revised abstract – items in red are added to meet the checklist requirements

Title: "Not just another walking program": Everyday Activity Supports You (EASY) model – a randomized pilot study for a parallel randomized controlled trial

Background: Maintaining physical activity is an important goal with positive health benefits, yet many people spend most of their day sitting. Our Everyday Activity Supports You (EASY) model aims to encourage movement through daily activities and utilitarian walking. The primary objective of this pilot trial was to test study feasibility (recruitment and retention rates) for the EASY model.
Methods: This 6-month parallel two-arm pilot trial took place in Vancouver, Canada (May to December 2013). Participants were healthy, inactive, community-dwelling women aged 55–70 years. We recruited through advertisements in local community newspapers and randomized participants using a remote web service. The model included: group-based education and social support, individualized physical activity prescription, and use of a Fitbit activity monitor. The control group received health-related information only. The main outcome measures were descriptions of study feasibility (recruitment and retention rates). We also collected information (blinded outcome assessment) on activity patterns, height and weight, blood pressure, and psychosocial variables.
Results: We advertised in local community newspapers to recruit participants. Over 3 weeks, 82 participants telephoned; following screening, 68% (56/82) met the inclusion criteria and 45% (25/56) were randomized by remote web-based allocation: 13 participants in the intervention group and 12 in the control group (education). At 6 months, 12/13 (92%; 95% CI 65% to 100%) intervention and 8/12 (67%; 95% CI 35% to 90%) control participants completed the final assessment. This met our a priori recruitment and retention criteria for success. Of those who declined 21/30 gave reasons of timing of sessions within working hours. There were no adverse events related to study participation.
Conclusions: The EASY pilot study was feasible to deliver; there was an increase in physical activity and reduction in weight and blood pressure for intervention participants at 6 months. In a future definitive trial greater drop-out in the control arm may be reduced by using a different design and alternative sources of recruitment might be considered.
Trial registration: ClinicalTrials.gov identifier: NCT01842061
Trial funding: Canadian Institute of Health Research, Michael Smith Foundation, Australian National Health and Medical Research Council.

Word count: 340

Item	Extension for pilot trials	Reported
Title	Identification of study as randomised pilot trial	✓
Trial design	Description of pilot trial design (e.g. parallel, cluster)	✓
METHODS		
Participants	Eligibility criteria for participants and the settings where the pilot trial was conducted	✓
Interventions	*Interventions intended for each group*	✓
Objective	Specific objectives of the pilot trial	✓
Outcome	Pre-specified assessment or measurement to address the pilot trial objective(s)[1]	✓
Randomisation	*How participants were allocated to interventions*	✓
Blinding (masking)	*Whether or not participants, care givers, and those assessing the objectives were blinded to group assignment*	✓
RESULTS		
Numbers randomised	Number of participants screened and randomised to each group for the pilot trial objective(s)[1]	✓
Recruitment	*Trial status*[2]	*N/A*
Numbers analysed	Number of participants analysed in each group for the pilot objective(s)[1]	✓
Outcome	Results for the pilot objective(s); including any expressions of uncertainty[1]	✓
Harms	*Important adverse events or side-effects*	✓
Conclusions	General interpretation of the results of pilot trial and their implications for the future definitive trial	✓
Trial registration	Registration number for pilot trial and name of trial register	✓
Funding	Source of funding for pilot trial	✓

Items above in italics are pilot trial checklist items unchanged from the standard RCT checklist.
[1] Space permitting, list all pilot trial objectives and give the results for each. Otherwise, report those which are a priori agreed as the most important (main) to the decision to proceed with the future definitive trial. [2] For conference abstracts.

Fig 2 | Revised version of example abstract for report of pilot trial,[21] shown alongside CONSORT for abstracts extension for pilot randomised trials

Title: "Not just another walking program": Everyday Activity Supports You (EASY) model – a randomized pilot study for a parallel randomized controlled trial

Background: Maintaining physical activity is an important goal with positive health benefits, yet many people spend most of their day sitting. Our Everyday Activity Supports You (EASY) model aims to encourage movement through daily activities and utilitarian walking. The primary objective of this pilot trial ~~phase~~ was to test study feasibility (recruitment and retention rates) for the EASY model.

Methods: This 6-month parallel two-arm pilot trial ~~study~~ took place in Vancouver, Canada, ~~from~~ (May to December 2013)~~, with data analyses in February 2014~~. Participants were healthy, inactive, community-dwelling women aged 55–70 years. We recruited through advertisements in local community newspapers and randomized participants using a remote web service. The model included ~~the following~~: group-based education and social support, individualized physical activity prescription ~~(called Activity 4-1-1)~~, and use of a Fitbit activity monitor. The control group received health-related information only. The main outcome measures were descriptions of study feasibility (recruitment and retention rates). We also collected information (blinded outcome assessment) on activity patterns~~ (ActiGraph GT3X+ accelerometers), and health-related outcomes such as body composition~~ (height and weight ~~using standard techniques~~), blood pressure ~~(automatic blood pressure monitor)~~, and psychosocial variables ~~(questionnaires)~~.

Results: We advertised in local community newspapers to recruit participants. Over 3 weeks, 82 participants telephoned; following screening, 68% (56/82) met the inclusion criteria and 45% (25/56) were randomized by remote web-based allocation: ~~This included~~ 13 participants in the intervention group and 12 ~~participants~~ in the control group (education). At 6 months, 12/13 (92%; 95% CI 65% to 100%) intervention and 8/12 (67%; 95% CI 35% to 90%) control participants completed the final assessment. This met our a priori recruitment and retention criteria for success. Of those who declined 21/30 gave reasons of timing of sessions within working hours. ~~Controlling for baseline values, the intervention group had an average of 2,080 [95% confidence intervals (CIs) 704, 4,918] more steps/day at 6 months compared with the control group. There was an average between group difference in weight loss of 4.3 [95% CI 6.22, 2.40] kg and reduction in diastolic blood pressure of 8.54 [95% CI 16.89, 0.198] mmHg, in favor of EASY.~~ There were no adverse events related to study participation.

Conclusions: The EASY pilot study was feasible to deliver; there was an increase in physical activity and reduction in weight and blood pressure for intervention participants at 6 months. In a future definitive trial greater drop-out in the control arm may be reduced by using a different design and alternative sources of recruitment might be considered.

Trial registration: ClinicalTrials.gov identifier: NCT01842061

Trial funding: Canadian Institute of Health Research, Michael Smith Foundation, Australian National Health and Medical Research Council.

Fig 3 | Track changes version of example abstract for report of pilot trial,[21] showing changes between figures 1 and 2

- Item 2b

- *Standard CONSORT item*: specific objectives or hypotheses

- *Extension for pilot trials*: specific objectives or research questions for pilot trial

- *Example 1 (listing objectives as primary and secondary)*

"In this feasibility trial, the research aim was to explore trial design, staff and resident acceptability of the interventions and outcome measures and to provide data to estimate the parameters required to design a definitive RCT . . . The primary objectives of the trial were as follows:

1. To assess how many care homes accepted the invitation to participate in research.

2. To determine whether the eligibility criteria for care home residents were too open or too restrictive by estimating feasible eligibility and recruitment rate.

3. To assess retention of care homes and residents by estimating 3 and 6-month follow-up rates.

4. To investigate the acceptability of nutritional support interventions to malnourished care home residents in terms of compliance and to care home staff in terms of adherence to the intervention schedule.

5. To assess the acceptability and feasibility (and factors influencing this) of the outcome measures as methods to measure efficacy of the interventions within a definitive trial.

The secondary objectives of the trial were as follows:

1. To investigate the completion of screening tools and questionnaires by care home staff.

2. To determine how many malnourished residents were able to participate in PROMs and to complete the questionnaires.

3. To pilot a Healthcare resource usage (HCRU) questionnaire.

4. To measure key outcome domains (for completion rates, missing data, estimates, variances and 95% confidence intervals for the difference between the intervention arms) for malnourished care home residents, including physical outcome measures and PROMs.

5. To collect and synthesise data, from which the Intracluster Correlation Coefficient (ICC) and sample size of a definitive cluster RCT (CRCT) could be estimated."[32]

- *Example 2 (objectives leading to a mixed methods study)*

"The main aim of the study is to assess the feasibility of conducting a definitive trial in terms of recruitment, use and acceptability of the intervention, follow-up at 3 and 6 months, and data collection methods. In addition, the study aims to establish suitable procedures for delivering the intervention and conducting assessments and procedures for ensuring recruitment and retention in the study. Finally, the study aims to discover whether using a structured, individualized approach to lifestyle assessment and referral will improve uptake and participation in lifestyle- and behaviour-change interventions.

The study will also examine, qualitatively, the acceptability of the assessment tool to patients in an acute

cardiology setting as well as patients' experiences of making lifestyle changes in order to develop effective recruitment and retention strategies.

The study will have a number of quantitative objectives:

1. To determine how many patients accept referral to a formal lifestyle programme;

2. To determine how many patients participate in a lifestyle-change intervention or initiate self-managed change;

3. To investigate the uptake of lifestyle intervention in relation to subsequent behaviour change and impact on health-related quality of life, mood and social satisfaction;

4. To estimate feasible eligibility, recruitment and refusal rates, and 3- and 6-months follow-up rates;

5. To measure key outcome domains (that is, for completion rates, missing data, estimates, variances and 95% confidence intervals for the difference between the control and intervention groups) for patients including clinical indicators and patient-reported measures of social satisfaction; health-related quality of life; and mood;

6. To synthesize data to inform the sample size of a definitive trial;

7. To determine the acceptability (and factors influencing this) of financial incentives as a method to encourage behaviour change, their pricing and factors influencing this."[39]

- *Explanation*

Although many aspects of feasibility may be related to each other, an articulation of specific objectives enables readers to understand the main areas of uncertainty to be addressed in the pilot trial and provides a working structure for presenting the methods and results in relation to these objectives. In addition, a comprehensive list of objectives enables other researchers to learn from and adopt similar approaches in their own studies.

It might be beneficial to separate the objectives into primary objectives (often those on which decisions about progressing to a future definitive RCT may be made) and secondary objectives, as in example 1, where feasibility objectives are primary and questions related to patient centred outcomes are treated as secondary. Because it is not always necessary to collect data on patient centred outcomes, it is important to give the rationale for collecting such data. For example, the purpose may be to ensure that certain data can be collected, including from specific patient groups (eg, elderly people, as in example 1), or to ensure that difficult-to-measure concepts such as lifestyle behaviour change can be assessed appropriately in the future definitive RCT (example 2). It might also be informative to state explicitly which objectives will be answered using quantitative methods and which using qualitative methods, as in example 2.

In example 2 the list of quantitative objectives are quite informative, but they are taken from the published study protocol. In the published pilot trial the objectives contained far less detail: "The Healthy Hospital Trial is a single-center, randomized controlled, 2-arm, parallel-group, unblinded feasibility trial that was conducted on 2 cardiology wards at the Leeds Teaching Hospitals Trust. Its primary aim was to explore the feasibility of individualized lifestyle referral assessment, estimate the rate of recruitment, and explore the feasibility of collecting the data and follow-up of participants to inform the sample size of a definitive trial. A secondary aim was to test the concept that an individually tailored assessment improves uptake of lifestyle change compared with usual assessment. The trial protocol has been published elsewhere."[40] We recommend putting detailed individual objectives into the pilot trial report itself so that readers can more easily judge the extent to which these have been fulfilled by the study.

Inclusion of an objective to test a hypothesis of effectiveness (or efficacy) is not recommended (see box 1). However, other kinds of hypotheses may be tested, such as when using an interim or surrogate outcome to address potential effectiveness.[41] (See also the section entitled Scope of this paper). However, a trial should always be adequately powered for any hypothesis test, and in a pilot trial it should be clearly stated that the objective is to assess potential effectiveness. If tests are carried out without adequate power (as they sometimes are in reality), they should certainly be viewed as secondary and a caveat included in the discussion.[21]

Methods

- Item 3a

- *Standard CONSORT item*: description of trial design (such as parallel, factorial) including allocation ratio

- *Extension for pilot trials*: description of pilot trial design (such as parallel, factorial) including allocation ratio

- *Example*

"We conducted a parallel-group randomised controlled pilot trial... An unequal randomisation of 2:1 vs 1:1 was chosen to provide experience delivering the hydration intervention to more patients."[42]

- *Explanation*

The design of any study should be described, be it a definitive trial or a pilot trial. It is not uncommon for pilot trials to adopt ratios other than the usual 1:1 for randomisation. 1:1 randomisation provides the greatest power for testing effectiveness in, for example, a future definitive RCT. However, a pilot trial commonly involves new, not established, interventions and one of the aims might then be to gain experience in delivering the intervention, in which case it is often better to have as many participants receiving the intervention as is feasible.

- Item 3b

- *Standard CONSORT item*: important changes to methods after trial commencement (such as eligibility criteria), with reasons

- *Extension for pilot trials*: important changes to methods after pilot trial commencement (such as eligibility criteria), with reasons

- *Example*

"After randomly assigning 11 patients (5 to standard care), we recognized that patients assigned to standard

care were receiving early surgery because, having achieved accelerated medical clearance, they were put on the operating room list. We therefore amended the protocol to randomly assign patients immediately on diagnosis; only those assigned to early surgery received an expedited medical assessment."[43]

- *Explanation*

Pilot trials are exploratory and so those conducting them should be able to modify the methods if a potential problem becomes apparent. In the case of Buse et al,[43] the original protocol specified that patients had to have medical clearance for rapid surgery before randomisation, but this led to contamination of the control group as some patients in this group were put on the surgical list for rapid surgery (accelerated surgery was the intervention) because it had been ascertained that they were suitable candidates. In the revised protocol participants were randomised first and then assessed for suitability to accelerated access. Thus the pilot potentially improved the design of the trial that was to follow. It is important to document all changes and give reasons for the changes. The example describes changes to the timing of randomisation, but there might also be changes to other aspects of the trial, such as the treatment regimen, eligibility criteria, or outcome variables.

- Item 4a

- *Standard CONSORT item*: eligibility criteria for participants

- *Example*

"Thirty-one sequential eligible people with HD [Huntington's disease] were recruited from the specialist HD clinics in Cardiff, the United Kingdom, and Oxford, the United Kingdom, between March 2011 and November 2011. Inclusion criteria were (1) diagnosis of HD, confirmed by genetic testing and neurological examination, (2) ability to walk independently as primary means of mobility, (3) willing to travel to the exercise center for the intervention, (4) capacity to give informed consent, (5) Unified Huntington's Disease Rating Scale Total Motor Score (UHDRS-TMS) and Total Functional Capacity (TFC) of at least 5/124 and 5/13, respectively, from last clinic visit, and (6) maintenance of a stable medical regimen for 4 weeks prior to initiation of study and considered by the recruiting clinician as able to maintain a stable regimen for the course of the study. Participants were not eligible if they (1) had a history of additional prior major neurological condition such as stroke, (2) had an orthopedic condition that limited mobility, (3) demonstrated uncontrolled psychiatric symptoms, (4) were pregnant, (5) demonstrated any contraindication to exercise, or (6) were involved in any interventional trial or within 3months of completing an interventional trial."[44]

- *Explanation*

Readers might want to know how the results of the trial are likely to apply to the future definitive RCT and other future trials with similar participants in similar settings. A variety of participants (eg, patients, doctors, assessors, caregivers, managers) might provide data to address objectives. For example, in a study in nursing homes, residents were interviewed to seek views on acceptability of the intervention, whereas nurses participated in focus groups to elicit views on randomisation or adherence to treatment protocol.[32] Eligibility criteria should be specified for each set of participants included in a pilot trial. The details provided must be specific enough to identify the clinical population and any other populations and the setting from which they were recruited and to confirm that legal issues were complied with, such as having capacity to give informed consent. Details should be sufficient to allow other researchers to interpret, learn from, and use the information provided.

- Item 4b

- *Standard CONSORT item*: settings and locations where the data were collected

- *Example*

"High-risk adolescents were recruited from three sources: (1) a sample of 205 offspring of BP parents between 12 and 18 years of age enrolled in the NIMH-funded Bipolar Offspring Study at the University of Pittsburgh (BIOS, PI: Birmaher); (2) offspring of adults receiving treatment for BP at Western Psychiatric Institute and Clinic (WPIC); and (3) siblings of youth receiving treatment for BP at the Child and Adolescent Bipolar Services clinic (CABS) at WPIC."[45]

- *Explanation*

The settings for recruiting patients and collecting data must be specified so that readers can judge the applicability (generalisability) of the findings to other trials as well as to the future definitive RCT. Authors should also make clear whether any pilot sites have particular features—for example, organisational features, characteristics that predispose the site to early adoption of new schemes, or specific relationships with the authors that could affect recruitment, consent, and follow-up. This is because these features may not be replicable in other sites and hence in future trials. As with item 4a, details must be sufficient to allow other researchers to interpret, learn from, and use the information.

- Item 4c

- *Extension for pilot trials*: how participants were identified and consented

- *Example*

"Between May and October 2013, clinical staff at participating gastroenterology outpatient clinics scanned and identified potential participants that met the study inclusion criteria. Then, either study invitation packs were sent to patients with researchers' contact details or patients seen consecutively in clinics were approached with the study information. All study information was co-designed with patients from the patient-involvement group. Interested participants then registered their interest with the researcher by telephone or email. This was followed up with a screening visit with the researcher and then informed written consent was obtained."[46]

- *Explanation*

This is a new item. It is especially important to report details of identification and consent in a pilot trial to allow the feasibility of the recruitment methods to be assessed. The way participants are identified and approached should be described in detail (eg, by adver-

doi: 10.1136/bmj.i5239 | *BMJ* 2016;355:i5239 | the bmj

tisement, or selection from medical records or another dataset) to enable readers to understand the generalisability (applicability) of the results. This might be of particular importance for scaling-up for the future definitive RCT, as well as being informative for other future trials. In addition, it is important to know of any specific aspects that might not be easy to implement in the future definitive RCT. Furthermore, a view is sometimes held that pilot trials do not need to be as rigorous in their processes as other trials, so it might be particularly important in these trials to show rigorous and ethical identification and recruitment processes. If details of the way participants were identified and consent obtained are already published in a protocol, then this should be clearly referenced.

- Item 5

- *Standard CONSORT item*: the interventions for each group with sufficient details to allow replication, including how and when they were actually administered

- *Example*

"Intervention (EXERcise or STRETCHing)

The amount of time required for participating in the exercise activities was the same for the EXER group and the STRETCH group. The only difference was the amount of energy expended during the activity. At the first session, the exercise trainer explained the procedures for the respective intervention (EXER or STRETCH), showed them the equipment available for the exercise or stretch sessions, and the coordinator familiarized the participant with the Actical device. The first two weeks required a minimum of 3 sessions at CI [Cooper Institute] for the trainer to teach them how to use the equipment and complete the exercise or stretch routines. Following the first 2 weeks, participants began doing their exercise program at home or other location (gym, park, etc.), and only had to come to CI once a week for an exercise session. Each EXER/STRETCH session averaged about 30-40 min.

EXERcise Intervention

Supervised exercise sessions at the Cooper Institute (CI) for the participants began by using the treadmills or stationary cycles. The CI trainers also taught patients how to complete home-based exercise sessions (e.g., choice of Wii Sports and Fit, jazzercise, jogging, weight training based on their preferred exercise) that were unsupervised workouts at the patient's home or in the community. The duration of each session generally was the time required to reach 1/3 or 1/4 of the total weekly caloric expenditure. There was a progression to the assigned exercise dose in the first few weeks that got them up to their minimum of 12 kilocalories/kilogram/week (KKW) energy expenditure (e.g., 8 KKW first week, 10 KKW second and 12 KKW by the third week). Participants exercised three times per week.

STRETCH Intervention

The stretch group spent approximately the same amount of time, but at energy expenditures of less than 4 KKW per session. After two weeks of three sessions at CI they moved to once a week at CI and two home-based sessions. A 5-10 minute stretching warm-up period included stretches that exercise the major muscle groups of the body. The series included such traditional "warm-up" stretches as: stretches of the gluts, inner thigh, calves and ankles, Achilles tendon, hamstring stretches, shoulder rolls forward and back, shoulder shrugs, isometrics for the neck hugging knees into the chest, moving forehead to right knee, then to left, then to both, and use of the pelvic tilt. An additional 10-15 minutes consisted of moving on to right and left calf stretches, quad stretches, and then to a series for the arms, hands, fingers, wrist, biceps/triceps, shoulders and back. All of the exercises were designed to be done slowly, emphasizing proper alignment, and rest periods to minimize overall physical exertion while obtaining general flexibility, and most importantly controlling for contact time with trainers and any social facilitation from participating in such activities. We had a different set of low level/low intensity routines for each of the 12 weeks to minimize boredom with the routines."[47]

- *Explanation*

If the pilot trial is to inform future research, the authors should report exact details of the treatment given to all study groups, and if one group receives treatment as usual this should also be described thoroughly. Details should include who administered the treatment, as well as what it comprised and how often and where it was delivered. The template for intervention description and replication (TIDieR) guidelines should be followed and the checklist completed.[48] If there are changes to the details of the treatments for any group, these must be reported (see item 3b).

- Item 6a

- *Standard CONSORT item*: completely defined prespecified primary and secondary outcome measures, including how and when they were assessed

- *Extension for pilot trials*: completely defined prespecified assessments or measurements to address each pilot trial objective specified in item 2b, including how and when they were assessed

- *Example*

"*Acceptability* and *demand* were assessed in terms of the usage and repeated usage of the intervention by the patients in the trial indicated by logged user statistics. The interventions' *practicability* was considered as the ability to log in and occurrence of constraints in delivery and was assessed in terms of the percentage of users in adolescents and professionals, its bounce percentage (percentage of login-errors) and other login-problems. The bounce-percentage was logged and participants were asked to report login-errors. *Integration* was assessed in terms of the extent to which our web-based intervention promotes care that was consistent with recognized standards of diabetes care for adolescents including those published by the International Diabetes Federation (IDF) in collaboration with the International Society for Pediatric and Adolescent Diabetes (ISPAD) and the American Diabetes Association (ADA; 3, 33); see also Appendix 1."[49]

- *Explanation*

In a definitive trial investigators are primarily interested in response variables or outcomes that enable

them to fulfil the primary objective (to assess the effect of an intervention or treatment), and a clear articulation of prespecified outcomes is required to guard against bias in the assessment of this effect. In a pilot trial, however, objectives should relate to feasibility (see box 1 and item 2b) and any measurements or assessments should enable these objectives to be addressed. To ensure the pilot trial meets its objectives, measures or assessments should be defined to address each separate objective or research question. In the example, objectives were to assess acceptability, demand, practicability, and integration. The authors list the measures used for each of these.

Variables that might be considered primary and secondary outcomes for the future definitive RCT might be measured in a pilot trial to assess response, completeness, or validity. The appropriate measures or assessments would then be response rates, completion rates, or measures of validity. Sometimes investigators may want to measure surrogate outcomes (see example in item 7b), variables on the causal pathway of what might eventually be the primary outcome in the future definitive RCT, or outcomes at early time points, in order to assess the potential for the intervention to affect likely outcomes in the future definitive RCT (see item 2b).

- Item 6b

- *Standard CONSORT item*: any changes to trial outcomes after the trial commenced, with reasons

- *Extension for pilot trials*: any changes to pilot trial assessments or measurements after the pilot trial commenced, with reasons

- *Example 1 (change to assessment time period)*

"Our outcome measures examined uptake and cessation because we hoped that our intervention would affect uptake by referring more people and the success rate of those referred by supporting adherence to treatment...The intervention had two distinct phases so, although not planned in the protocol, we examined uptake of services and 4-week quit rates by trial arm, in these two periods."[50]

- *Example 2 (change to measurement instrument)*

"We defined . . . initiation of change as participation in a formal program or a self-directed program that was intended to result in change either in diet, physical activity, smoking, or alcohol consumption at any time (binary) . . . In our published protocol, we had proposed 4 categories of change, but we found it difficult to distinguish between "persisted" and "maintained" in the qualitative follow-up interviews; hence, we combined persistence and maintenance of change in 1 category."[40]

- *Explanation*

An assessment or measure might change during a pilot trial because the change enables investigators to glean more information about the operation of the intervention (as in example 1) or for reasons of acceptability or practicability (example 2). In example 2 it became impractical to use a measurement instrument with four categories when it was identified that researchers could not distinguish between two of the categories. In the interests of full reporting and because of the usefulness of such information to others working in the same specialty, all such changes should be reported.

- Item 6c

- *Extension for pilot trials*: if applicable, prespecified criteria used to judge whether, or how, to proceed with future definitive trial

- *Example*

"Feasibility (delivery) and acceptability (uptake) of the DECISION+ program were the main outcome measures of this pilot trial. Investigators had established a priori threshold for specific feasibility and acceptability criteria. These were the following: (a) the proportion of contacted FMGs [Family medicine groups] participating in the pilot study would be 50% or greater, (b) the proportion of recruited family physicians participating in all three workshops would be 70% or greater, (c) the mean level of satisfaction from family physicians regarding the workshops would be 65% or greater, and (d) the proportion of missing data in each completed questionnaire would be less than 10%."[34]

- *Explanation*

This is a new item. The purpose of a pilot trial is to assess the feasibility of proceeding to the next stage in the research process. To do this investigators need some criteria on which to base the decision about whether or not to proceed. The next stage in the research process will normally, although not always, be the future definitive RCT.

The UK National Institute for Health Research requires that pilot or feasibility studies have clear criteria for deciding whether or not to progress to the next stage: "We expect that when pilot or feasibility studies are proposed by applicants, or specified in commissioning briefs, a clear route of progression criteria to the substantive study will be described. Listing clear progression criteria will apply whether the brief or proposal describes just the preliminary study or both together. Whether preliminary and main studies are funded together or separately may be decided on practical grounds."[51]

In many pilot studies, however, such criteria may be best viewed as guidelines rather than strict thresholds that determine progression. In the example, the authors found that only 24% of the family medicine groups (FMGs) agreed to participate. They state "Not reaching the pre-established criteria does not necessarily indicate unfeasibility of the trial but rather underlines changes to be made to the protocol".[34] Clearly it is important to discuss whether such changes to protocol are likely to be feasible, and this discussion might often benefit from input independent of the trial team—for example, from the trial steering committee. This would be a reason for having such a committee in place for a pilot trial. Bugge et al recently provided further guidance on decision making after a pilot trial.[52]

In addition to the possibility of making changes to the trial protocol, investigators should also be aware that estimates of rates in pilot trials may be subject to considerable uncertainty, so that it is best to be cautious about setting definitive thresholds that could be missed simply due to chance variation.[41] In fact it is

doi: 10.1136/bmj.i5239 | *BMJ* 2016;355:i5239 | the bmj

becoming increasingly common for investigators to use a traffic light system for criteria used to judge feasibility, whereby measures (eg, recruitment rates) below a lower threshold indicate that the trial is not feasible, above a higher threshold that it is feasible, and between the two that it might be feasible if appropriate changes can be made.

- Item 7a

- *Standard CONSORT item*: how sample size was determined

- *Extension for pilot trials*: rationale for numbers in the pilot trial

- *Example 1 (rationale based on assessment of practicalities and estimating rates)*

"Since this was a pilot study, a sample size calculation was not performed. The researchers aimed for 120 participants because it was felt this would be a large enough sample to inform them about the practicalities of delivering several self- management courses led by patients with COPD, recruitment, uptake, and attrition."[53]

- *Example 2 (rationale based on percentage of number required for future definitive RCT)*

"As this is a feasibility study a formal sample size calculation is not required, but we estimated the number of participants required as around 10% of the number required for the Phase 3 trial. The sample size calculation for the Phase 3 trial suggests we need to recruit 1665 participants. Given the participant population, a high level of attrition may be anticipated. We therefore aim to recruit 200 participants to the feasibility trial to inform the design and sample size of the Phase 3 RCT."[54]

- *Explanation*

The criterion of congruency between the objectives and the sample size holds as true for a pilot trial as for any study. Many pilot trials have key objectives related to estimating rates of acceptance, recruitment, retention, or uptake (see item 2b for examples). For these sorts of objectives, numbers required in the study should ideally be set to ensure a desired degree of precision around the estimated rate, although in practice it may be difficult to achieve these numbers. Additionally, for pilot trials where the key objective focuses on the acceptability or feasibility of introducing the intervention, it might be useful to consider how many sites are needed, as the acceptability or feasibility of introduction can sometimes depend on the site. In example 1, the authors state their reason for choosing their required sample size in relation to estimating rates and to exploring practicalities of implementing the intervention. They could, however, have provided stronger justification for their chosen number, such as likely recruitment or attrition rate and desired precision around these rates, so that the reader (and funder) has more grounds for believing the trial could achieve its objectives beyond a feeling.

Most methodological papers that focus on recommendations about sample size requirements for pilot trials assume that the main aim of such a trial is to estimate a quantitative measure such as the variance (or standard deviation) of an effect size to inform the sample size calculation for a future definitive RCT. Methods focus on the precision with which such estimates can be obtained. There are several relevant papers.[55 56 57] Among these, Whitehead et al suggests that the size of a pilot trial should be related to the size of the future definitive RCT.[58] For such a trial designed with 90% power and two sided 5% significance, they recommend pilot trial sample sizes for each treatment arm of 75, 25, 15, and 10 for standardised effect sizes that are extra small (0.1), small (0.2), medium (0.5), or large (0.8), respectively.

Example 2 illustrates another approach that uses a sample that is a certain percentage of the expected size of the future definitive RCT. The authors reference the paper by Cocks and Torgerson, which is based on using a sample size under which a one sided 80% confidence interval for the effect size will exclude the minimum clinically important difference if the null hypothesis is true.[59] This is a similar calculation to that used in estimating sample size needed for efficacy or effectiveness but allows for additional uncertainty in the resulting effect size estimate, thus effectively assessing potential effectiveness. If an objective is to assess potential effectiveness using a surrogate or interim outcome, investigators will need to use a standard sample size calculation to ensure there is adequate power. However, this type of objective is rare in pilot trials.

- Item 7b

- *Standard CONSORT item*: when applicable, explanation of any interim analyses and stopping guidelines

- *Example*

"The board members were instructed to perform an interim analysis after 60 patients had been enrolled, at which point they could recommend stopping the trial if an overwhelming effect was detected on the basis of the critical significance level (P≤0.02), as adjusted for the Lan–DeMets alpha-spending function with Pocock boundary"[20]

- *Explanation*

As pilot trials are small, it is uncommon for them to define criteria for early stopping, but if they do, these should be reported. The example is a pilot trial testing a surrogate outcome. There was considerable uncertainty about the variability of this outcome measure, and so the authors calculated a conservative sample size but included an interim analysis after recruiting 60 patients, in case their a priori estimates were too large and they had enough information at that stage to inform subsequent trials.

- Item 8a

- *Standard CONSORT item*: method used to generate the random allocation sequence

- *Example*

"Participants were randomly allocated to the intervention 'MBCT group' or 'wait-list control group' . . . Random allocation was computer generated."[46]

- *Explanation*

Randomisation induces unpredictability in the allocation of each unit of randomisation. This is an important element of ensuring an unbiased treatment effect in RCTs evaluating effectiveness or efficacy because in the

long run it ensures balance in characteristics between intervention groups. In a pilot trial, the soundness of the randomisation method might not directly influence robustness of the pilot trial results, which are not focused on estimates of effectiveness or efficacy, but a clear description of the process of randomisation is still important for transparent reporting.

In addition, in some pilot trials one of the objectives might be to assess the feasibility of randomisation; it is also important, therefore, that details are reported. If assessing feasibility involves more than one method being used to generate a random allocation sequence, each method should be described adequately.

- Item 8b

- *Standard CONSORT item*: type of randomisation; details of any restriction (such as blocking and block size)

- *Extension for pilot trials*: type of randomisation(s); details of any restriction (such as blocking and block size)

- *Example 1 (example with blocking)*

"Participants were randomised in block sizes of three by computer-generated randomisation to the hydration group or the control group (2:1), stratified by gender."[42]

- *Example 2 (two different types of randomisation)*

"In addition to random allocation to one of the three treatment arms, we used a 2 × 2 factorial design to distribute practices and participants across two trial design factors: cluster versus individual allocation and systematic versus opportunistic recruitment (see Fig 1). We randomly assigned 24 practices (8 practices in each of 3 geographical regions (Bristol, Devon and Coventry)) in a 3:1 ratio to cluster (practice) allocation or individual allocation, and in a 1:1 ratio to opportunistic or systematic recruitment. The differential allocation ratio with regard to randomisation method was due to the need to ensure even numbers of practices and participants in each of the three arms across the cluster randomised practices."[60]

- *Explanation*

The type of randomisation, including whether simple or restricted, should be reported.

For practical reasons simple randomisation is sometimes used in pilot trials even when restricted randomisation is expected to be used in the future definitive RCT, and if this is the case this needs to be described.

Restricted randomisation is particularly useful in small trials evaluating the effectiveness of an intervention, to ensure balance in certain characteristics between intervention and control groups (see main CONSORT statement).[2 61] In pilot trials, restricted randomisation might be used to mimic the type of randomisation expected in the future definitive RCT or, if it is deemed important, to have balanced groups even if restricted randomisation is not expected to be used in the future definitive RCT. In example 1, stratified randomisation, employing blocking, was used.

One of the objectives of a pilot trial might be to assess the feasibility of randomisation; it is therefore possible that different types of randomisation could be tried, as in example 2 where cluster versus individual randomisation was considered.[60]

- Item 9

- *Standard CONSORT item*: mechanism used to implement the random allocation sequence (such as sequentially numbered containers), describing any steps taken to conceal the sequence until interventions were assigned

- *Example*

"Allocation . . . was implemented using an automated telephone randomization service provided by the Bristol Randomized Trials Collaboration to ensure concealment from clinical staff undertaking recruitment."[62]

- *Explanation*

Ensuring allocation concealment is a cornerstone of a good randomised trial design. This mechanism performs a key function in minimising bias by preventing foreknowledge of treatment assignment, which could influence those who enrol participants. In a future definitive RCT a single mechanism will be used to conceal allocation. However, in a pilot trial the main purpose of using an allocation concealment mechanism is to establish the feasibility of the mechanism. If there is considerable uncertainty about the mechanism to be used, more than one mechanism may be tried in the pilot trial. We would expect this to be rare, but when it does occur the details of each mechanism tried should be fully described.

- Item 10

- *Standard CONSORT item*: who generated the random allocation sequence, who enrolled participants, and who assigned participants to interventions

- *Example 1 (who generated the random allocation sequence)*

"An independent statistical consultant set up the web-based randomization process to assign eligible participants to intervention or control groups by remote allocation, using permuted blocks of sizes 2 and 4. No one directly involved in the project had access to allocation codes."[21]

Example 2 (who enrolled participants, who assigned participants to interventions)

"Eligible children and their families were identified by the clinician conducting the assessment. If the child and his or her family were willing to find out more about the study a researcher contacted the family and arranged to visit them at a convenient location (usually at home) . . . Those willing to take part were randomized to receive either specialist medical care or to specialist medical care plus the Phil Parker Lightning Process (LP). Allocation . . . was implemented . . . by the Bristol Randomized Trials Collaboration . . . "[62]

- *Explanation*

It is important that the pilot trial confirms that allocation concealment can be implemented in a way that could be replicated in the future definitive RCT. This involves knowing who generated the randomisation sequence and who enrolled and assigned participants.

- Item 11a

- *Standard CONSORT item*: if done, who was blinded after assignment to interventions (eg, participants, care providers, those assessing outcomes) and how

doi: 10.1136/bmj.i5239 | *BMJ* 2016;355:i5239 | the bmj

- *Example 1 (blinding of multiple people)*

"Patients, families, ICU [intensive care unit] staff, ultrasound technologists, and research personnel were all blinded to drug allocation. The study pharmacist at each center was the only person who was not blinded."[63]

- *Example 2 (placebo controlled)*

"A synbiotic formulation (Synbiotic 2000®) containing 4 strains of probiotic bacteria (10^{10} each) plus 4 non-digestible, fermentable dietary fibers (2.5 g each) was provided each day, versus a fiber-only placebo formulation."[64]

- *Explanation*

In the future definitive RCT investigators will want to reduce the chance of a biased result as much as possible. Blinding is seen as one of the most effective ways of doing this, at least in trials where blinding is feasible (see main CONSORT statement for details). The main purpose of a pilot trial is to assess the feasibility of methods, including those to reduce bias. In some pilot trials it might be useful to report the method of blinding in detail, as in example 2, to help readers who might want to replicate the method in future RCTs.

It is tempting in a pilot trial to try and assess the success of blinding by asking people whether they believed they were blinded or not. This was done, for example in Arnold et al.[65] This is not recommended, however, because evidence suggests that results of doing this largely reflects the effectiveness of the intervention rather than anything else.[66]

- Item 11b

- *Standard CONSORT item*: if relevant, description of the similarity of interventions

- *Example*

"Each study drug infusion was administered using a standard volume-based rate escalation protocol preceded by the administration of 100 mg of hydrocortisone intravenously, 50 mg of diphenhydramine orally or intravenously, and 650 mg of acetaminophen orally to minimize infusion-related reactions and avoid unblinding."[65]

- *Explanation*

If blinding is done by creating a placebo, it is important in trials assessing the effect of an intervention to detail what features of the placebo were made similar to the active intervention (usually a drug)—for example, appearance, taste, smell, method of being administered. However, many of the interventions described in pilot trials are not drug interventions. Nevertheless, it remains important to describe what was done to try and ensure that the intervention and control arms received identical treatment aside from the active ingredient where this is possible. It is equally important to note that for complex interventions it might not be possible or feasible to blind certain people to allocation using these types of methods.

- Item 12a

- *Standard CONSORT item*: statistical methods used to compare groups for primary and secondary outcomes

- *Extension for pilot trials*: methods used to address each pilot trial objective whether qualitative or quantitative

- *Example 1 (descriptive and narrative reporting)*

"The feasibility outcomes were reported descriptively and narratively. For the clinical endpoints, only descriptive statistics, mean (standard deviation) for continuous outcomes and raw count (%) for categorical outcomes, were reported."[67]

- *Example 2 (confidence intervals)*

"For the primary outcomes, the feasibility criteria were the recruitment rate and duration, retention rate, safety, adverse events, compliance, acceptability of the interventions and fatigue . . . The recruitment rate, consisting of the eligibility and consent rate, was calculated with 95% CI . . . Medians (range) were reported for ordinal data (fatigue), mean (95% confidence interval (CI)) were reported for continuous data (walking speed and walking distance) and raw count (number, %) was reported for nominal data. Due to the nature of this feasibility study, it was decided not to conduct any efficacy statistical tests on the walking and fatigue data."[68]

- *Explanation*

A range of methods can be used to address the objectives in a pilot trial. These need not be statistical. Providing information about the methods used ensures that findings can be verified on the basis of the description of the analyses used. The primary focus is on methods for dealing with feasibility objectives. These methods are often based on descriptive statistics such as means and percentages but might also be narrative descriptions (example 1). Typically, any estimates of effect using participant outcomes as they are likely to be measured in the future definitive RCT would be reported as estimates with 95% confidence intervals without P values—because pilot trials are not powered for testing hypotheses about effectiveness.

- Item 12b

- *Standard CONSORT item*: methods for additional analyses, such as subgroup analyses and adjusted analyses

- *Extension for pilot trials*: not applicable

- *Explanation*

In a definitive trial, analyses of a difference in treatment effect for subgroups or analysis of outcomes adjusted for baseline imbalance might provide useful information. However, such analyses in a pilot trial are not applicable because the primary focus is not on determining treatment effects or differences in effects between subgroups. Rather, the focus is on assessing feasibility or piloting procedures to inform the design of the future definitive RCT.

Results

- Item 13a

- *Standard CONSORT item*: for each group, the numbers of participants who were randomly assigned, received intended treatment, and were analysed for the primary outcome

- *Extension for pilot trials*: for each group, the numbers of participants who were approached and/or assessed for eligibility, randomly assigned, received intended treatment, and were assessed for each objective

- *Example*

See figures 4 and 5.[2 69]

- *Explanation*

As for other trials, we recommend a diagram for communicating the flow of participants in a pilot trial. A flow diagram is a key element of the CONSORT statement and has been widely adopted.[70] A review of RCTs published in five leading general and internal medicine journals found that reporting was considerably more thorough in articles that included a diagram of the flow of participants through a trial, as recommended by CONSORT.[70] A complete CONSORT flow diagram also reduces the time for readers to find essential information to assess the reliability of a trial. It is also likely to improve the availability of some information that otherwise might not be reported.

Information required to complete a CONSORT flow diagram includes the number of participants evaluated for potential enrolment into the trial and the numbers of participants who were randomly assigned to each intervention group, received treatment as allocated, completed treatment as allocated, and were analysed for the primary outcome, with numbers and reasons for exclusions at each step.[2 61]

For pilot trials it might also be important to know the number of participants who were approached (or screened) before being assessed for eligibility for potential enrolment into the trial. This ensures that readers can assess external validity and how representative the trial participants are likely to be compared with all eligible participants.[71] Additionally, for pilot trials it is important to know how many participants were approached before being evaluated for potential enrolment in the trial and how easy it was to recruit them, in order to assess the potential for enrolment for the future definitive RCT and other future trials. In some cases where these elements are a major focus of a pilot trial more information may be needed in the flow diagram (fig 4).

For pilot trials it is appropriate to report the number of participants assessed for each pilot trial objective, rather than the number analysed for the primary outcome (as would be the case for the future definitive RCT). If there are a limited number of objectives in the pilot trial then all should be listed and results for each objective reported in the flow diagram. If there are multiple objectives, then agreement should be reached a priori about which are the most important to decide whether to proceed to a future definitive RCT, and only these objectives should be reported in the flow diagram. Figure 5 provides a template for a CONSORT flow diagram for pilot trials, including presentation of results for different objectives. The exact form and content might, however, vary in relation to the specific features of the trial. Authors should ensure that their flow diagram matches the key objectives as far as possible.

- Item 13b

- *Standard CONSORT item*: for each group, losses and exclusions after randomisation, together with reasons

- *Example*

"All 16 patients randomised to the Symptoms Clinic attended the first appointment and 11 completed either three or four appointments. Of the remainder, two were clearly improving at the time they were seen and agreed to early discharge; two found further attendance difficult after a second appointment and one declined any further contact after the first appointment. Several patients randomised to usual care expressed some disappointment at the time of their allocation, although follow-up response rates were comparable between the two groups."[72]

- *Explanation*

For some RCTs the flow of participants through each phase of the trial can be relatively straightforward to describe, particularly if there were no losses to follow-up or exclusions. However, in more complex trials, it might be difficult for readers to identify whether and why some participants did not receive the treatment as allocated, were lost to follow-up, or were excluded.[73] In a definitive trial this information is crucial for interpreting generalisability, as participants who are excluded after allocation are unlikely to be representative of all participants in the study.[74] In a pilot trial, this information could be used to judge potential generalisability of the future definitive RCT but also to assess the acceptability of an intervention to participants and to aid planning of the future definitive RCT and other trials in similar settings.

- Item 14a

- *Standard CONSORT item*: dates defining the periods of recruitment and follow-up

- *Example*

"Patient enrolment started in August 2003 and was completed in October 2005."[75]

- *Explanation*

It is important to report dates for all studies for transparency. An added rationale for pilot trials is that factors such as disease definitions, treatment options, and reimbursement plans that could affect the future definitive RCT might have changed between the date that the pilot trial was conducted and the date the future definitive RCT starts. The availability of different treatments outside the trial can also change and might make a difference to people's willingness to be randomised. Thus recruitment to a pilot trial could be easier, or more difficult, than recruitment to the future definitive RCT. In addition, knowing the length of time over which the study took place might be important for planning the future definitive, and other, RCTs.

- Item 14b

- *Standard CONSORT item*: why the trial ended or was stopped

- *Extension for pilot trials*: why the pilot trial ended or was stopped

- *Example 1 (stopped without reaching intended recruitment but provided sufficient data)*

"Enterotoxigenic Escherichia coli (ETEC) is a major cause of travellers' diarrhoea . . . We designed this phase II, double-blind, randomised placebo-controlled study to investigate the epidemiology of natural infection with ETEC in placebo recipients with a planned enrolment of 300 individuals, at a placebo-to-LT patch ratio of 2:1 . . . The study was halted when enrolment

doi: 10.1136/bmj.i5239 | *BMJ* 2016;355:i5239 | the bmj

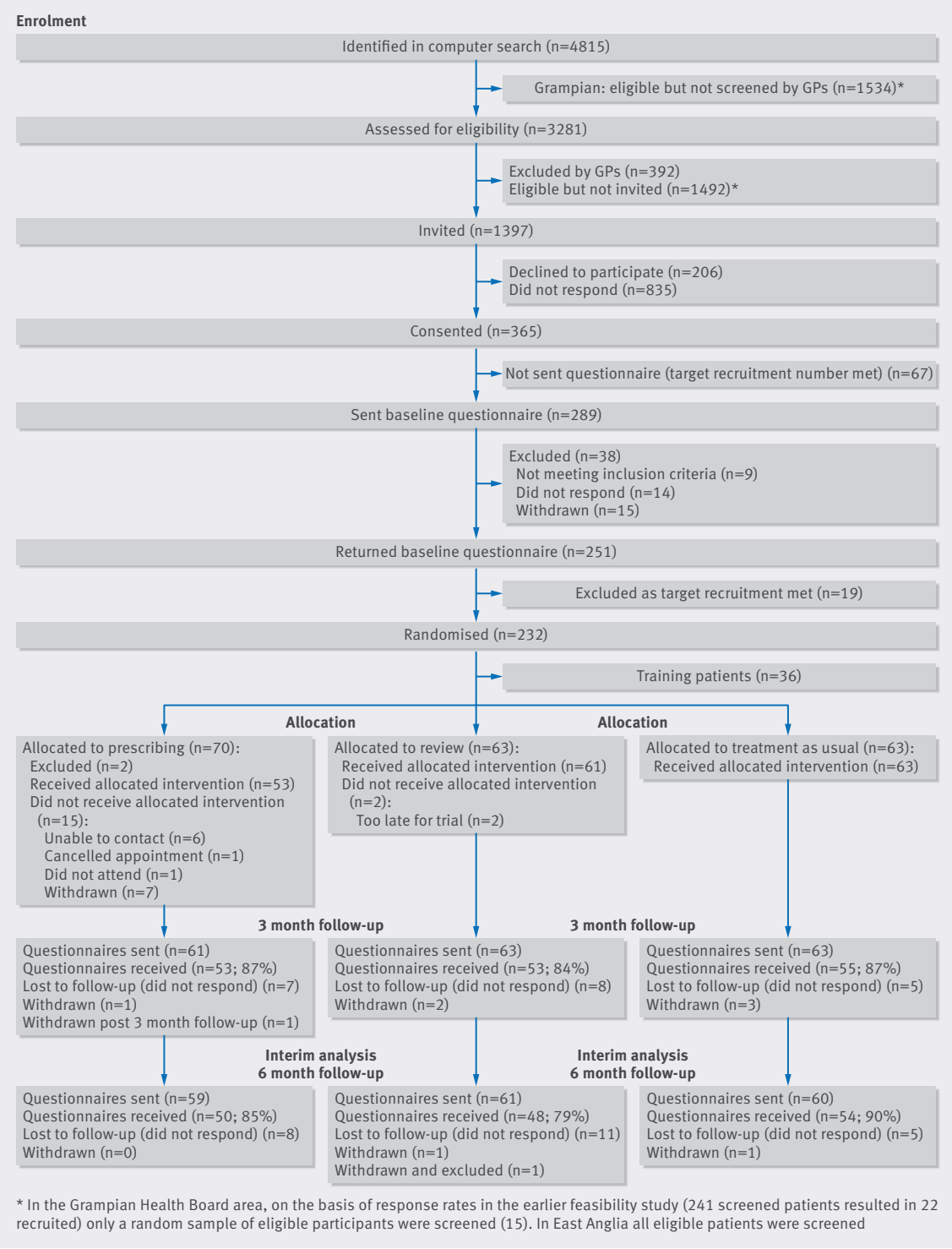


Fig 4 | **Flow diagram of a randomised pilot trial of pharmacist led management of chronic pain in primary care (reproduced from Bruhn et al[69])**

reached 201, because the planned interval for conduct had been exceeded, and it was thought that a placebo group greater than 100, although less powerful than the original 200, would be sufficient to assess the ETEC attack rate in placebo recipients . . . 24 (22%) of 111 placebo recipients had diarrhoea, of whom 11 (10%) had ETEC diarrhoea."[76]

- Example 2 (stopped at end of recruitment but did not provide sufficient data)

"Recruitment rates were lower than expected which led to the study being expanded to further areas and opened to self-referral via advertisement. However, because of better management of hypertension due to changes in the UK Quality and Outcomes Framework

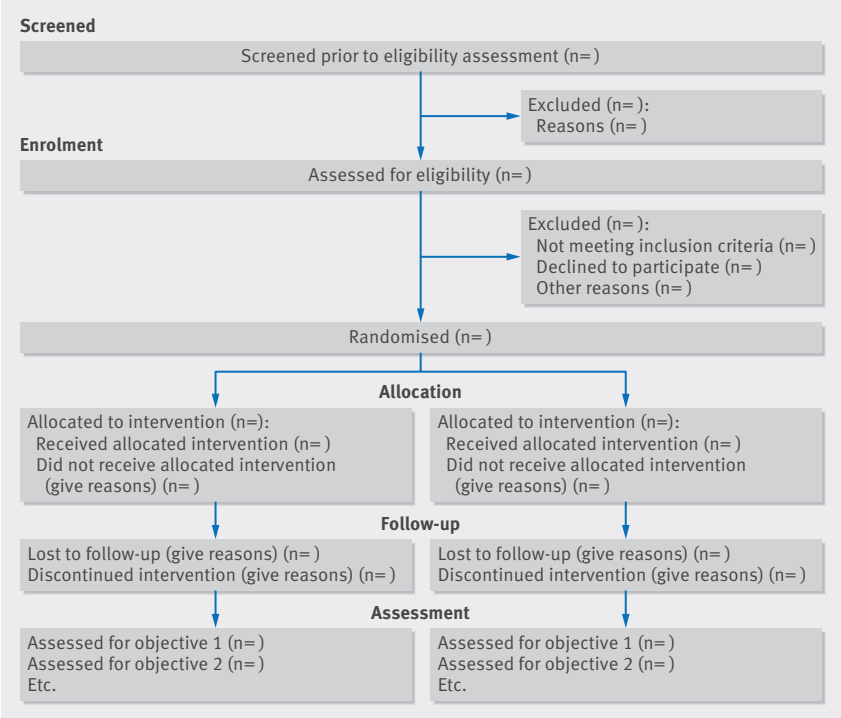


Fig 5 | **Recommended flow diagram of progress through phases of a parallel randomised pilot trial of two groups—that is, screening, enrolment, intervention allocation, follow-up, and assessment for each pilot trial objective. Adapted from Moher et al[2]**

guidelines for blood pressure treatment, few eligible patients were identified and the study closed at the end of the recruitment period, with 13 participants consenting, but 12 failing screening resulting in one recruited participant."[77]

- *Explanation*

When pilot trials end or are stopped, it is important to state why as this might affect the feasibility of the future definitive RCT. In example 1 the investigators had run out of time and thought they would have sufficient participants to estimate the rate of diarrhoea so as to inform future studies. It is not uncommon for changes in the clinical environment to occur, leading to fewer patients with unmanaged disease, and this can lead to major studies, not just pilot studies, failing to recruit. This illustrates a benefit of a pilot study to assess the likely accrual for a future definitive RCT. In example 2 the reason for stopping was simply a failure to recruit, and the reasons for this are clearly stated. Other potential reasons for stopping include the intervention being impossible to implement, other studies indicating that the research has become irrelevant, and difficulties with funding. It is also helpful to know who made the decision to stop early. In definitive RCTs a data monitoring committee often makes recommendations to stop the trial. It might not be necessary to have data monitoring committees for all pilot trials, but investigators should give some thought as to how the decision to stop should be made.

- Item 15

- *Standard CONSORT item*: a table showing baseline demographic and clinical characteristics for each group

- *Example*

- *Explanation*

In an RCT evaluating the effect of an intervention, a table of baseline characteristics is important to indicate any differences between intervention groups that could affect the face validity of the trial. In a pilot trial, the number of participants is likely to be smaller than in the future definitive RCT and baseline imbalances might therefore be more likely. Similar to a definitive trial, imbalance does not suggest bias, and in any case bias is not a problem in the same way it is in a definitive trial because an assessment of the effect of an intervention is not the primary concern. Nevertheless, baseline data are important to aid interpretation of the results, including a consideration of generalisability, and a table is the best way of presenting this information.

- Item 16

Standard CONSORT item: for each group, number of participants (denominator) included in each analysis and whether the analysis was by original assigned groups

- *Extension for pilot trials*: for each objective, number of participants (denominator) included in each analysis. If relevant, these numbers should be by randomised group

- *Example 1 (number of sites contacted)*

"A research assistant made 41 introductory phone calls to contact the medical directors of the 21 eligible FMGs [family medicine groups] over a four-week period. One director could not be contacted. Information leaflets were faxed to the 20 contacted FMGs."[34]

doi: 10.1136/bmj.i5239 | *BMJ* 2016;355:i5239 | the bmj

- *Example 2 (number of practitioners taking part within sites)*

"Out of the 52 eligible family physicians working in the five participating FMGs [family medicine groups], 39 (75%) agreed to participate in the study."[34]

- *Explanation*

In RCTs evaluating the effect of an intervention, outcomes are usually measured on participants and therefore denominators are numbers of participants. However, because of the potential variety of objectives in a pilot trial, the denominators for measures that assess feasibility according to these objectives might be organisations, health practitioners, patients, or, in some cases, episodes or events. In the interests of simplicity we have not changed the word "participants" in this item, but the item should be interpreted in the light of the particular objective and associated measure or assessment. The two examples are taken from the same trial. One objective was to assess the feasibility of recruitment. Participants for that objective are FMGs (example 1) and family physicians (example 2). The denominators of 21 (FMGs) and 52 (family physicians) indicate numbers approached and therefore the effort involved in recruiting. In this example providing numbers by randomised group is not relevant.

- Item 17a

- *Standard CONSORT item*: for each primary and secondary outcome, results for each group, and the estimated effect size and its precision (such as 95% confidence interval)

- *Extension for pilot trials*: for each objective, results including expressions of uncertainty (such as 95% confidence interval) for any estimates. If relevant, these results should be by randomised group

- *Example 1 (feasibility outcome)*

"The ABSORB [A bioabsorbable everolimus-eluting coronary stent system for patients with single de-novo coronary artery lesions] study aimed to assess the feasibility and safety of the BVS [bioasorbable everolimus-eluting stent] stent in patients with single de-novo coronary artery lesions . . .Procedural success was 100% (30/30 patients), and device success 94% (29/31 attempts at implantation of the stent)."[78]

- *Example 2 (proposed outcome in future definitive trial)*

"Rates of initiation of lifestyle change also favoured the individualized assessment arm but less clearly. At 3 months, 75% of the individualized assessment arm and 68% of the usual assessment arm had initiated changes in their lifestyle (unadjusted odds ratio, 1.38 [95%CI, 0.55 to 3.52]). At 6 months, the percentages were 85% and 75%, suggesting increased initiation of change over time in both arms, with the gap widening slightly (unadjusted odds ratio, 1.86 [95% CI, 0.64 to 5.77]) . . . Wide CIs again point to the degree of uncertainty around this conclusion"[40]

- *Explanation*

It is important that the reported results of a pilot trial reflect the objectives. Results might include, for example, recruitment, retention or response rates, or other sorts of rates, as in example 1. Because the sample size in a pilot trial is likely to be small, estimates of these rates will be imprecise and this imprecision should be recognised, for example, by calculating a confidence interval around the estimate. Commonly, authors do not give such a confidence interval, but if the numerator and denominator are given the confidence interval can be calculated. In example 1 the Wilson 95% confidence interval for 100% (30/30) is 88.65% to 100% and for 94% (29/31) is 79.78% to 98.21% (OpenEpi Seattle).[78] If authors do report differences between trial arms (and this is not necessary if it is not consistent with the objectives of the trial) then confidence intervals again provide readers with an assessment of precision (example 2), which usually indicates considerable uncertainty. If samples in the pilot trial and future definitive RCT are drawn from slightly different populations, confidence intervals calculated from the pilot will not directly indicate the likely upper and lower bounds of the relevant measure in the future definitive RCT, but can nevertheless highlight the lack of precision effectively.

- Item 17b

- *Standard CONSORT item*: for binary outcomes, presentation of both absolute and relative effect sizes is recommended

- *Extension for pilot trials*: not applicable

- *Explanation*

This item is included in the 2010 CONSORT statement because when considering clinical implications, neither the relative nor the absolute measures of effect size for binary outcomes give a complete picture of the effect of an intervention. For example, relative risks are less affected by differences in baseline populations across studies than are absolute risks, although sometimes can be misinterpreted in terms of population benefit. In addition, different audiences (clinical, policy, patient) prefer to use one or the other measure. However, in pilot trials the situation is different. Because of the imprecision of estimates from these trials and the fact that samples in these trials can be unrepresentative (see item 17a), we caution against any reliance on estimates of effect size from pilot trials for clinical implications (see also Introduction, Scope of this paper, and box 1). Information from outcome data, however, can be legitimately used for other purposes, such as estimating inputs for sample size for the future definitive RCT (see item 7a). Thus item 17b, which is underpinned by rationale around clinical implications, is not applicable.

- Item 18

- *Standard CONSORT item*: results of any other analyses performed, including subgroup analyses and adjusted analyses, distinguishing prespecified from exploratory

- *Extension for pilot trials:* results of any other analyses performed that could be used to inform the future definitive RCT

- *Example*

"Sensitivity analysis

At both six and 12 weeks, findings were insensitive to the exclusion of those catheterised throughout their

hospital stay (and also to the exclusion of those who were never incontinent following the removal of a catheter). However, at both time points, odds ratios reduced when those with pre-stroke incontinence were excluded . . . "[79]

- *Explanation*

It is possible that the results of analyses that were not initially planned might have important implications for the future definitive RCT. Such findings should be reported and discussed in relation to how they might inform the future definitive RCT. In the example, although numbers were small, the authors inferred from the unplanned sensitivity analyses that those with pre-stroke incontinence were at least as likely, or more likely, to benefit from the intervention than those continent pre-stroke, and concluded that this group of patients should be included in the full trial.

- Item 19

- *Standard CONSORT item*: all important harms or unintended effects in each group (for specific guidance see CONSORT for harms)[7]

- *Example 1 (potential harm)*

"Intervention and usual treatment groups were similar in terms of age, gender, and marital status, but those in the intervention group were more likely to be unemployed (69% v. 59%), to use methods other than poisoning (23% v. 9%), to have a past history of self-harm (67% v. 53%) and to have had previous psychiatric treatment (64% v. 53%).

Online Table DS1 shows self-harm repetition and resource use in the two groups. The 12-month repeat rate for individuals in the intervention group was 34.4% v. 12.5% for the usual treatment group (odds ratio (OR) 3.67, 95% CI 1.0–13.1 . . .) . . . Adjusting for baseline clinical factors (centre, method of harm (self-poisoning v. other), previous self-harm, previous psychiatric treatment), the odds ratio for repetition and incidence rate ratio for number of repeat episodes remained elevated . . . "[80]

- *Example 2 (unintended effect or potential harm)*

"An unanticipated finding in this study was a 4-kg weight loss, on average, favouring the intervention group, although we recognized that there were some differences in weight between groups at study commencement that may have had an effect on our results . . . Thus, there is a clear role for dietary considerations in any study that aims to positively influence body weight. Although we provided one educational session on nutrition during a tour of a local grocery store with a dietitian and modelled healthy food choices with the lunches provided, dietary behaviors and body weight were not the focus of the study."[21]

- *Explanation*

It is crucial to report all important or potential harms or unintended effects on individual participants in each group to enable the study design for the future definitive RCT to be changed either to avoid these effects or to put in place effective processes for monitoring potential harms. In example 1, it was not clear whether the unexpected increased risk of repeated self harm in the intervention group was real or a consequence of baseline covariate imbalance, or peculiar to the particular setting. This led to a proposal to change the design to use stratified randomisation in the future definitive RCT. In example 2, the unintended effect of weight loss in elderly participants led to the decision to include a dietary component in the intervention to avoid potential harm in the future definitive RCT. This information might also be useful to other researchers planning similar studies.

- Item 19a

- *Extension for pilot trials*: if relevant, other important unintended consequences

- *Example (unintended consequence)*

"Twelve of the 13 active, and 11 of the 13 traditional practices recruited a total of 231 participants in the 12 months from mid-April 1998. Active practices recruited 165 (average practice recruitment rate of 1.71 per 1000 registered patients, i.e., 141% of expected) while traditional practices recruited only 66 (0.57 per 1000 i.e. 54% of expected) (Fig 1). On average active practices recruited 12.7 participants (range 0-39), while traditional practices recruited only 5.1 participants (range 0-18) (Table 2). Although both types of practices recruited similar percentages of those identified (13% in active; 16% in traditional), active identified 1257, far more than the 416 by traditional practices. The extreme difference in recruitment rates led to an investigation of baseline characteristics of participants in the two groups (Table 3). Participants recruited by active practices were more likely to be working full-time and to have had further education since leaving school. They were also suffering from milder back pain, less limited physically and less depressed."[81]

- *Explanation*

This is a new item reflecting the importance of reporting unintended consequences that do not directly affect individual participants but might have implications for the validity of the future definitive RCT if not dealt with in the pilot trial. By unintended consequences we mean things that happened in the pilot trial that the investigators did not intend to look for but that would have such implications. In the example, the design of the pilot trial included practice level randomisation, with participant recruitment after that randomisation. This had unintended consequences in the balance of recruited participants between arms, and in the main study the researchers abandoned randomisation at the practice level.

Discussion

- Item 20

- *Standard CONSORT item*: trial limitations, addressing sources of potential bias, imprecision, and, if relevant, multiplicity of analyses

- *Extension for pilot trials*: pilot trial limitations, addressing sources of potential bias and remaining uncertainty about feasibility

- *Example 1 (pilot trial limitations)*

"In some cases, platelet mass was calculated on an MPV [mean platelet volume] that was up to 72 hours old based on our previous research on the relationship between platelet mass and IVH [intraventricular hemor-

doi: 10.1136/bmj.i5239 | *BMJ* 2016;355:i5239 | the bmj

Table 3 | Example of baseline information for each group. From Seebacher et al[68]

Parameter	Group A Music cued motor imagery (n=10)	Group B Metronome cued motor imagery (n=10)	Group C Control group (n=10)
Females to males	10:0	7:3	5:5
Age (years)[a]	47.3 (38.4, 56.2)	41.8 (34.8, 48.8)	46.1 (39.8, 52.5)
EDSS[b]	3 (1.5, 4.5)	2.5 (1.5, 4.5)	2.5 (1.5, 4.0)
MFIS total score[b]	35 (3, 67)	32 (17, 50)	33.5 (0, 48)
Participants with fatigue (MFIS total score ≥38)	4/10	2/10	4/10
T25FW (s)[a]	6.1 (4.5, 7.6)	5.4 (4.5, 6.2)	5.2 (4.3, 6.1)
6MWT (m)[a]	453.1 (365.0, 541.1)	428.2 (352.8, 503.6)	484.7 (399.5, 569.8)

EDSS Expanded Disability Status Scale, MFIS Modified Fatigue Impact Scale, T25FW Timed 25-Foot Walk, s seconds, 6MWT 6-Minute Walk Test, m metres.
aMean (95% confidence interval).
bMedian (range).

rhage]. We cannot rule out the possibility that during acute thrombocytopenia changes in MPV may be more acute. Because platelet counts were not confirmed by manual count, we cannot exclude the unlikely possibility that some infants may have had pseudothrombocytopenia."[82]

- *Example 2 (potential bias)*

"Fourth, the house staff at the two academic centers in the study may have been a source of contamination. Additional house staff occasionally provided overnight coverage at the intervention group academic center. These additional house staff were not formally educated about the study, so they effectively functioned as if they were in the control group. Conversely, additional house staff who provided overnight coverage at the control group academic center may have been previously educated about our study while working at the intervention group academic center. Thus, they effectively functioned as if they were in the intervention group."[83]

- *Example 3 (remaining uncertainty)*

"The integration of a nested, internal pilot in the definitive trial should also be considered to allow continued monitoring of the feasibility, in particular, the assessment of using different inclusion criteria and the recommended changes to the data collection methods, particularly within the first year of recruitment. The use of a qualitative element to assess the participants' views on data collection methods would also be beneficial."[84]

- *Explanation*

Identifying and discussing the limitations of a study helps to provide a better context for understanding the importance of its findings. In a pilot trial it might also be helpful to distinguish between limitations that can be overcome in a future definitive RCT, and those that cannot. In example 1 the authors explain the limitation of a method of measurement although they do not say whether they think this could be overcome in a future definitive RCT.

In a future definitive RCT, investigators will want, as far as possible, to avoid sources of bias that might affect treatment effect estimates. In a pilot trial, investigators are not primarily interested in treatment effect, so these biases will not be of so much concern but it would still be useful to identify potential biases that could affect the treatment effect in the future definitive RCT so that investigators have a better chance of avoiding these. In example 2 a potential source of bias in the future definitive RCT is identified.

If substantial areas of uncertainty about feasibility remain at the end of the pilot trial that prevent investigators from proceeding with a future definitive RCT or warrant investigation in an internal pilot then, for clarity, these should be reported, as in example 3.

Lastly, although we do not recommend this, if underpowered tests are performed and reported then investigators should always point out this limitation to avoid misinterpretation of results (see item 2b).

- Item 21

- *Standard CONSORT item*: generalisability (external validity, applicability) of the trial findings

- *Extension for pilot trials*: generalisability (applicability) of pilot trial methods and findings to future definitive trial and other studies

- *Example 1 (generalisability of findings)*

"We accommodated variability in choice and duration of standard treatments to enhance generalizability of the results and had high rates of follow-up."[65]

- *Example 2 (generalisability to other pilot trials)*

"Our data reflect the activities of only one pilot trial; however, we hope that the methods may serve as a template for analyzing other pilot studies with different designs in other settings."[63]

- *Example 3 (generalisability concerns)*

"Although safety issues must remain paramount in practice and clinical research, common overstringent exclusion criteria may increase perceived trial safety yet limit the generalizability of trial results and delay answers to important clinical questions. Reevaluation of the PROTECT Pilot exclusion criteria will . . . enhance the applicability of the larger PROTECT study . . . The PROTECT Pilot indicated the need for another pilot study (DIRECT) to determine the safety of dalteparin 5000 IU SC OD among patients with severe renal insufficiency (creatinine clearance, b30 mL/min)."[63]

- *Explanation*

Generalisability (applicability) is the extent to which aspects of a study can be applied to other circumstances. Generalisability is not absolute and is a matter of judgment. In a definitive trial, readers are usually interested in the generalisability of findings to situations outside research settings—for example, routine clinical practice. However, in pilot trials this is not the case because the size of these studies does not allow this. Nevertheless, it might be important to consider generalisability at the pilot stage as this could be important for the generalisability of the future definitive RCT (example 1), the findings and the methods might be applied in research settings other than the future definitive RCT (example 2), or there might be concerns about the generalisability of results from a future definitive RCT conducted in an identical way to the pilot trial that might lead to changes in the design of the future definitive RCT or further piloting (example 3).

- Item 22

- *Standard CONSORT item*: interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence

- *Extension for pilot trials*: interpretation consistent with pilot trial objectives and findings, balancing potential benefits and harms, and considering other relevant evidence

- *Example 1 (consistency with objectives and findings)*

"One of the goals of this pilot study was to investigate the feasibility of using platelet transfusion guidelines based on platelet mass. In five infants, MPV [mean platelet volume] was not available within 72 hours preceding the diagnosis of thrombocytopenia. A lack of immediately available MPV may limit the clinical utility and generalizability of this transfusion strategy at some institutions...In our study approximately half of the families at the Christiana Hospital site did not consent to the study. This information is important for planning future studies on platelet transfusion. Many families were unable to decide on enrollment at a time when their infant was thrombocytopenic and facing transfusion. An alternative study design for platelet transfusion study may involve enrolling a larger number of infants on admission, regardless of platelet count, with transfusion guidelines to apply only if they actually become thrombocytopenic. This approach may limit the stress on families of being approached about the need for transfusion and a transfusion related study simultaneously."[82]

- *Example 2 (considering other relevant evidence)*

"As far as we know, our participants were able to perform motor imagery. Our results seem to be in contrast to previous studies demonstrating a lower capacity for motor imagery in people with MS. However, these authors linked impaired motor imagery in this population particularly to cognitive dysfunction and depression. Therefore, persons with cognitive impairment and depression were excluded from our study. Several studies used patient-rated questionnaires, such as the Kinaesthetic and Visual Imagery Questionnaire to assess the motor imagery ability in their participants. Our study could have used this patient-rated questionnaire, but our participants were called weekly to ask for any problems with kinaesthetic motor imagery, and they were supported accordingly. In addition, all motor imagery ability studies in people with MS were experimental studies with no long-term training effects, in contrast to our 4 weeks duration study with 24 training sessions which might have enhanced the mental representation."[68]

- *Example 3 (consistency with findings in relation to decision criteria)*

"Moreover, the results of this trial support the feasibility and acceptability of conducting a large clustered randomised trial involving dyads of family physicians and their patients in SDM regarding the optimal use of antibiotics for ARI. This conclusion is reached even if not all predetermined standards for our criteria were always fully met. Indeed, it has been established that not reaching the preestablished criteria does not necessarily indicate unfeasibility of the trial but rather underlines changes to be made to the protocol . . . 24% of the eligible FMGs agreed to take part in the study, less than the 50% expected. We were probably too confident when targeting a 50% positive response rate from all identified FMGs."[34]

- *Explanation*

Interpretation of findings helps increase understanding of the importance of the results. In example 1, in addition to matching their interpretation to one of the goals of the study, the authors draw out the issue of redesign to reduce stress in families approached and so increase recruitment—and hopefully eventually a positive benefit for the children involved. This observation could be helpful to others planning similar studies. As for definitive trials, readers will want to know how the evidence presented in the report of a pilot trial relates to evidence from other sources (example 2). These sources might be other feasibility studies carried out by the authors or studies by different authors in the same or similar settings or with similar patients. If a priori decision criteria have been used (item 6c) then interpretation should be made with reference to these criteria (example 3).

- Item 22a

Extension for pilot trials: implications for progression from pilot to future definitive trial, including any proposed amendments

- *Example 1 (proposed amendments to improve recruitment)*

"The target of recruit to time was met but this did not translate to the expected number of eligible patients being recruited. Eligibility of the screened population was much lower than expected, indicating that the inclusion criteria may have been too stringent. The exclusion criteria of BMI ≤22 kg/m2 was based on published evidence that a BMI at the lower end of the normal range can increase mortality in the haemodialysis population . . . However, body composition is thought to play a much greater role in the protective effects of a greater BMI, than the BMI itself . . . The use of BMI as a screening tool was a quick and easy measure but the level of ≤22 kg/m2 should be reassessed prior to a definitive trial. If the BMI was raised to ≤24 kg/m2 then this would have increased potential recruitment by 10%."[84]

- *Example 2 (proposed amendments to improve cooperation)*

"Six homes declined to actively participate before even beginning the intervention. To ensure cooperation by the entire team and avoid early withdrawal, a short presentation to the Professional Advisory Committee team could potentially boost recruitment/retention. Obtaining initial consent from both the medical director and director of care may also be beneficial. Furthermore, to overcome logistical challenges, particularly for homes in the far north, providing an opportunity to view modules on a Web site or participate remotely may improve participation."[85]

- *Example 3 (implications for progression to future definitive RCT)*

doi: 10.1136/bmj.i5239 | *BMJ* 2016;355:i5239 | the bmj

"Hospitals that were allocated to receive our multicomponent intervention comprising education, standardized paper-based physician orders, and group audit and feedback did not have a higher rate of hospitalized medical patients appropriately managed for thromboprophylaxis within 24 hours of admission than did hospitals that were not allocated to this strategy (63% vs. 67%). This finding, coupled with the problems associated with ensuring preprinted orders were placed in all medical charts led us to conclude that this intervention should not be provided on a larger scale without major revision and testing. That is, it was not feasible."[83]

- *Explanation*

This is a new item. To progress from a pilot trial to a future definitive RCT, it is important to understand how the implications of the findings in the pilot carry over to the future definitive RCT. To aid clarity, a simple statement as to whether the future definitive RCT will be planned without any changes from the pilot trial, planned with changes from the pilot trial (examples 1 and 2), or not planned because of major problems with feasibility (example 3), is sufficient. If it is proposed to plan the future definitive RCT with specific changes from the pilot trial, these should be stated.

Other information

- Item 23

- *Standard CONSORT item*: registration number and name of trial registry

- *Extension for pilot trials*: registration number for pilot trial and name of trial registry

- *Example*

"Trial registration number: Clinical Trials, protocol registration system: NCT01695070."[86]

- *Explanation*

It is just as important for a pilot trial to be registered with a unique identifier as it is for a definitive trial. Registration ensures transparency and accountability and in the United Kingdom is now a requirement for all clinical trials before approval from UK ethics committees.[87 88] It ensures all ongoing work is in the public domain, and subsequent publication (and therefore access to findings for the greater good) confirmed. The World Health Organization states that "the registration of all interventional trials is a scientific, ethical and moral responsibility."[89] The International Committee of Medical Journal Editors requires all trials to be registered as a criterion for publication and lists suggested registries.[90]

- Item 24

- *Standard CONSORT item*: where the full trial protocol can be accessed, if available

- *Extension for pilot trials*: where the pilot trial protocol can be accessed, if available

- *Example 1 (reference to published protocol)*

"The Healthy Hospital Trial is a single-center, randomized controlled, 2-arm, parallel-group, unblinded feasibility trial that was conducted on 2 cardiology wards at the Leeds Teaching Hospitals Trust. Its primary aim was to explore the feasibility of individualized lifestyle referral assessment, estimate the rate of recruitment, and explore the feasibility of collecting the data and follow-up of participants to inform the sample size of a definitive trial. . . . The trial protocol has been published elsewhere."[40]

- *Example 2 (protocol as supporting information)*

"The protocol for this trial and supporting TREND checklist are available as supporting information; see Checklist S1 and Protocol S1."[91]

- *Example 3 (protocol available from authors on request)*

"Participants in the control arm (but not the other two arms) received a 16-page informational booklet relevant to education, medical care, housing, employment, and community resources (protocol available from authors upon request)."[92]

- *Explanation*

Access to the full protocol for the pilot trial is important as it will prespecify all the main components of the trial. The SPIRIT (standard protocol items: recommendations for interventional trials) statement defines an evidence based set of items that would be included.[93] Accessibility of the protocol allows subsequent output to be checked for completeness, and reduces the chance of selective reporting to suggest "better" results. The examples illustrate the different ways in which protocols may be made available, such as prior publication (example 1), as an addendum to the report of the pilot trial (example 2), or on request from the authors (example 3). Options where the protocol is already in the public domain, such as prior publication, are to be preferred. Other methods that could be used to achieve this would include publication on a study website. Trial registries (see item 23) also include some core protocol items.

- Item 25

- *Standard CONSORT item*: sources of funding and other support (such as supply of drugs), role of funders

- *Example*

Funding: "This trial was funded through grants from Academic Health Science Centres Alternative Funding Plan Innovation Fund of Ontario and Octapharma Canada. The trial funders had no role in the design of the study, the collection, analysis or interpretation of data, the writing of the report, or the decision to submit the article for publication."[94]

- *Explanation*

Reporting the sources of all funding for a pilot trial (that is, the main research award and any other support, such as supply of equipment) allows readers to judge the potential influence of the funding body on the design, conduct, analysis, and reporting of the trial. If no specific funding was provided to support the pilot trial, this should also be stated. As reported in the main CONSORT statement, a systematic review has shown that research funded by the pharmaceutical industry is more likely to report findings in its favour, compared with reports of research funded by independent funding bodies.[2 61] Where funders have had no involvement in any aspect of trial conduct or reporting this should be explicitly stated.

- Item 26

- *Extension for pilot trials*: ethical approval/research review committee approval confirmed with reference number

- Example

"The Regional Ethical Review Board at the Karolinska Institute approved the study, no. 2007/1401-31/3."[95]

- Explanation

This is a new item that has been added to the CONSORT checklist because of the need to emphasise that all research, including pilot trials, should only be conducted within an ethical framework and with all ethical and other approvals in place before commencement. Of particular relevance to pilot trials is the need also to be aware of any restrictions imposed by the reviewing ethical committee, because these would have implications for the design and conduct of the future definitive RCT.

Comment

Reports of RCTs need to include key information on the methods and results so that readers can accurately interpret the contents of the report. This is as true for pilot trials as it is for any other RCT. The CONSORT 2010 statement provides the latest recommendations from the CONSORT Group on essential items to be included in the report of an RCT.[2 61] However, pilot trials differ from other randomised trials in their aims and objectives, focusing on assessing feasibility rather than effectiveness or efficacy. Therefore, although much of the information to be reported in these trials is similar to that which needs to be reported in any other randomised trial, there are some key differences in the type of information and in the appropriate interpretation of standard CONSORT reporting items.

In this article we introduce and explain these key differences in an extension to the CONSORT checklist specific to pilot trials. In the section entitled "Scope of this paper" we discuss several other types of feasibility study, and "proof of concept" trials. Other researchers have begun to look at the transfer of ideas between these different types of study (eg, Wilson et al[96]). It is our expectation that some of the principles of reporting outlined in this extension can be adapted for other types of feasibility or proof of concept studies.

Use of the CONSORT statement for the reporting of two group parallel trials is associated with improved reporting quality.[97] We believe that the routine use of this proposed extension to the CONSORT statement will result in similar improvements in reporting of pilot trials. When reporting a pilot trial, authors should address each of the 26 items on the CONSORT extension checklist using this document, referring to the main CONSORT guidelines as appropriate. Adherence to the CONSORT statement and extensions can also help researchers designing trials in the future and can guide peer reviewers and editors in their evaluation of manuscripts. Many journals recommend adherence to the CONSORT recommendations in their instructions to authors. We encourage them to direct authors to this and to other extensions of CONSORT for specific trial designs. A tool is currently being developed to support journals in doing this.[98] The most up to date versions of all CONSORT recommendations are available at www.consort-statement.org.

PAFS consensus group authors (listed alphabetically): Doug Altman (address 1), Frank Bretz (2), Marion Campbell (3), Erik Cobo (4), Peter Craig (5), Peter Davidson (6), Trish Groves (7), Freedom Gumedze (8), Jenny Hewison (9), Allison Hirst (10), Pat Hoddinott (11), Sarah E Lamb (12), Tom Lang (13), Elaine McColl (14), Alicia O'Cathain (15), Daniel R Shanahan (16), Chris Sutton (17), Peter Tugwell (18)

1 Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK

2 Statistical Methodology and Consulting, Novartis Pharma AG, Basel, Switzerland, and Informatics and Intelligents Systems, Medical University of Vienna, Austria

3 Health Services Research Unit, University of Aberdeen, Foresterhill, Aberdeen, UK

4 Department of Statistics and Operations Research, UPC, Barcelona-Tech, Spain

5 MRC/CSO Social and Public Health Sciences Unit, University of Glasgow, Glasgow, UK

6 National Institute for Health Research, Evaluation, Trials, and Studies Coordinating Centre, University of Southampton, Southampton, UK

7 BMJ, BMA House, London, UK

8 Department of Statistical Sciences, University of Cape Town, Cape Town, South Africa

9 Leeds Institute of Health Sciences, University of Leeds, Leeds, UK

10 IDEAL Collaboration, Nuffield Department of Surgical Sciences, University of Oxford, Oxford, UK

11 Nursing Midwifery and Allied Health Professionals Research Unit, Faculty of Health Sciences and Sport, University of Stirling, Scotland, UK

12 Oxford Clinical Trials Research Unit, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK

13 Tom Lang Communications and Training International, Kirkland, Washington, USA

14 Institute of Health and Society, Newcastle University, Newcastle Upon Tyne, UK

15 School of Health and Related Research, University of Sheffield, Sheffield, UK

16 BioMed Central, London, UK

17 College of Health and Wellbeing, University of Central Lancashire, Preston, UK

18 Department of Medicine, University of Ottawa, Ottawa, Canada

During the development of this work we presented our thinking at workshops and open meetings and would like to thank all participants for their valuable input and views: Society for Clinical Trials May 2013 Boston; Clinical Trials Methodology Conference Edinburgh November 2013; Royal Statistical Society October 2013; UK National Institute for Health Research (NIHR) Research Design Service London and South East March 2014 London; UK NIHR statisticians April 2015 London; Society for Academic Primary Care Annual Scientific Meeting July 2015 Oxford; HSRPP Health Services Research and Pharmacy Practice Conference April 2016. We also thank Colin Begg for feedback on the proposed items to be included in the guidelines during the two day consensus meeting in Oxford.

Contributors: SME, MJC, CMB, SH, LT, and GAL conceived the study. SME led its development and execution. SME, CLC, MJC, CMB, SH, LT, and GAL contributed to various aspects of the empirical work, and the PAFS consensus group provided feedback on the proposed items to be included in the guidelines through a consensus meeting and by email. SME, CLC, MJC, CMB, SH, LT, and GAL drafted the manuscript and all authors reviewed it and approved the final version. SME is the guarantor.

Funding: We received grants from Queen Mary University of London (£7495), University of Sheffield (£8000), NIHR RDS London and South East (£2000), NIHR Statisticians Group (£2400), and Chief Scientist Office Scotland (£1000). CLC (nee Coleman) was funded by a National Institute for Health Research (NIHR) research methods fellowship. This article presents independent research partly funded by the NIHR. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health. Marion Campbell works at the Health Services Research Unit, University of Aberdeen, and the Unit receives core funding from the Scottish

Government Health and Social Care Directorates; however, the opinions expressed are those of the authors alone. The funders had no role in this study.

Competing interests: All authors have completed the ICMJE disclosure form at http://www.icmje.org/coi_disclosure.pdf and declare support from the following organisations for the submitted work—Queen Mary University of London, University of Sheffield (ScHARR Research Committee Pump priming grant), NIHR Research Design Services London and South East, NIHR Statisticians Group, Chief Scientist Office Scotland. GAL is editor in chief of the new BioMed central journal *Pilot and Feasibility Studies* proposed by Daniel R Shanahan, which was created out of this work. Daniel R Shanahan is employed by BioMed Central, and Trish Groves is an editor of the *BMJ*. SME, CMB, MJC, and LT are on the editorial board of *Pilot and Feasibility Studies* and CLC is an associate editor. Frank Bretz works for Novartis. The authors declare no other competing interests. None of the listed people involved in either journal played any part in the peer review process or editorial decision making.

Ethical approval: The Delphi study was approved by the ScHARR research ethics committee at the University of Sheffield.

Data sharing: Owing to a requirement by the ethics committee that the authors specified when the data from the Delphi study will be destroyed, the authors are not able to give unlimited access to the Delphi study quantitative data. These anonymised data are available from SME on request to all interested researchers up until the end of 2018. Qualitative data from the Delphi study are not available because these more sensitive data cannot be anonymised and there was no consent from participants to share potentially identifiable data.

Transparency: The lead author (SME) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

1 Schulz KF, Altman DG, Moher D. CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. *Ann Intern Med* 2010;152:726-32. doi:10.7326/0003-4819-152-11-201006010-00232.
2 Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010;340:c869. doi:10.1136/bmj.c869.
3 Turner L, Moher D, Shamseer L, et al. The influence of CONSORT on the quality of reporting of randomised controlled trials: an updated review. *Trials* 2011;12(Suppl 1):A47doi:10.1186/1745-6215-12-S1-A47.
4 Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJ. CONSORT Group. Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement. *JAMA* 2006;295:1152-60. doi:10.1001/jama.295.10.1152.
5 Campbell MK, Piaggio G, Elbourne DR, Altman DG. CONSORT Group. Consort 2010 statement: extension to cluster randomised trials. *BMJ* 2012;345:e5661. doi:10.1136/bmj.e5661.
6 Zwarenstein M, Treweek S, Gagnier JJ, et al. CONSORT group Pragmatic Trials in Healthcare (Practihc) group. Improving the reporting of pragmatic trials: an extension of the CONSORT statement. *BMJ* 2008;337:a2390. doi:10.1136/bmj.a2390.
7 Ioannidis JP, Evans SJ, Gøtzsche PC, et al. CONSORT Group. Better reporting of harms in randomized trials: an extension of the CONSORT statement. *Ann Intern Med* 2004;141:781-8. doi:10.7326/0003-4819-141-10-200411160-00009.
8 Boutron I, Moher D, Altman DG, Schulz KF, Ravaud P. CONSORT Group. Extending the CONSORT statement to randomized trials of nonpharmacologic treatment: explanation and elaboration. *Ann Intern Med* 2008;148:295-309. doi:10.7326/0003-4819-148-4-200802190-00008.
9 Gagnier JJ, Boon H, Rochon P, Moher D, Barnes J, Bombardier C. CONSORT Group. Reporting randomized, controlled trials of herbal interventions: an elaborated CONSORT statement. *Ann Intern Med* 2006;144:364-7. doi:10.7326/0003-4819-144-5-200603070-00013.
10 Calvert M, Blazeby J, Altman DG, Revicki DA, Moher D, Brundage MD. CONSORT PRO Group. Reporting of patient-reported outcomes in randomized trials: the CONSORT PRO extension. *JAMA* 2013;309:814-22. doi:10.1001/jama.2013.879.
11 Eldridge SM, Lancaster GA, Campbell MJ, et al. Defining feasibility and pilot studies in preparation for randomised controlled trials: development of a conceptual framework. *PLoS One* 2016;11:e0150205. doi:10.1371/journal.pone.0150205.
12 Shanyinde M, Pickering RM, Weatherall M. Questions asked and answered in pilot and feasibility randomized controlled trials. *BMC Med Res Methodol* 2011;11:117. doi:10.1186/1471-2288-11-117.
13 Lancaster GA, Dodd S, Williamson PR. Design and analysis of pilot studies: recommendations for good practice. *J Eval Clin Pract* 2004;10:307-12. doi:10.1111/j.2002.384.doc.x.
14 Thabane L, Ma J, Chu R, et al. A tutorial on pilot studies: the what, why and how. *BMC Med Res Methodol* 2010;10:1. doi:10.1186/1471-2288-10-1.
15 Arain M, Campbell MJ, Cooper CL, Lancaster GA. What is a pilot or feasibility study? A review of current practice and editorial policy. *BMC Med Res Methodol* 2010;10:67. doi:10.1186/1471-2288-10-67.
16 Cartwright ME, Cohen S, Fleishaker JC, et al. Proof of concept: a PhRMA position paper with recommendations for best practice. *Clin Pharmacol Ther* 2010;87:278-85. doi:10.1038/clpt.2009.286.
17 Fisch R, Jones I, Jones J, Kerman J, Rosenkranz GK, Schmidli H. Bayesian design of proof-of-concept trials. *Ther Innov Regul Sci* 2015;49:155-62. doi:10.1177/2168479014533970.
18 Stallard N. Optimal sample sizes for phase II clinical trials and pilot studies. *Stat Med* 2012;31:1031-42. doi:10.1002/sim.4357.
19 Burke DL, Billingham LJ, Girling AJ, Riley RD. Meta-analysis of randomized phase II trials to inform subsequent phase III decisions. *Trials* 2014;15:346. doi:10.1186/1745-6215-15-346.
20 Talmor D, Sarge T, Malhotra A, et al. Mechanical ventilation guided by esophageal pressure in acute lung injury. *N Engl J Med* 2008;359:2095-104. doi:10.1056/NEJMoa0708638.
21 Ashe MC, Winters M, Hoppmann CA, et al. "Not just another walking program": Everyday Activity Supports You (EASY) model-a randomized pilot study for a parallel randomized controlled trial. *Pilot Feasibility Stud* 2015;1:1-12. doi:10.1186/2055-5784-1-4.
22 Hoddinott P, Craig L, Maclennan G, Boyers D, Vale L. NHS Grampian and the University of Aberdeen FEST project team. Process evaluation for the FEeding Support Team (FEST) randomised controlled feasibility trial of proactive and reactive telephone support for breastfeeding women living in disadvantaged areas. *BMJ Open* 2012;2:e001039. doi:10.1136/bmjopen-2012-001039.
23 Schoultz M, Macaden L, Hubbard G. Participants' perspectives on mindfulness-based cognitive therapy for inflammatory bowel disease: a qualitative study nested within a pilot randomised controlled trial. *Pilot Feasibility Stud* 2016;2:3. doi:10.1186/s40814-015-0041-z.
24 O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. *Acad Med* 2014;89:1245-51. doi:10.1097/ACM.0000000000000388.
25 Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *Int J Qual Health Care* 2007;19:349-57. doi:10.1093/intqhc/mzm042.
26 O'Cathain A, Hoddinott P, Lewin S, et al. Maximising the impact of qualitative research in feasibility studies for randomised controlled trials: guidance for researchers. *Pilot Feasibility Stud* 2015;1:32. doi:10.1186/s40814-015-0026-y.
27 Stansfeld SA, Kerry S, Chandola T, et al. Pilot study of a cluster randomised trial of a guided e-learning health promotion intervention for managers based on management standards for the improvement of employee well-being and reduction of sickness absence: GEM Study. *BMJ Open* 2015;5:e007981. doi:10.1136/bmjopen-2015-007981.
28 O'Cathain A, Thomas KJ, Drabble SJ, Rudolph A, Hewison J. What can qualitative research do for randomised controlled trials? A systematic mapping review. *BMJ Open* 2013;3:e002889. doi:10.1136/bmjopen-2013-002889.
29 Hoddinott P, Craig L, Maclennan G, Boyers D, Vale L. NHS Grampian and the University of Aberdeen FEST Project Team. The FEeding Support Team (FEST) randomised, controlled feasibility trial of proactive and reactive telephone support for breastfeeding women living in disadvantaged areas. *BMJ Open* 2012;2:e000652. doi:10.1136/bmjopen-2011-000652.
30 Thabane L, Hopewell S, Lancaster GA, et al. Methods and processes for development of a CONSORT extension for reporting pilot randomized controlled trials. *Pilot Feasibility Stud* 2016;2:25. doi:10.1186/s40814-016-0065-z
31 Gilbody S, Peckham E, Man M-S, et al. Bespoke smoking cessation for people with severe mental ill health (SCIMITAR): a pilot randomised controlled trial. *Lancet Psychiatry* 2015;2:395-402. doi:10.1016/S2215-0366(15)00091-7.
32 Stow R, Ives N, Smith C, Rick C, Rushton A. A cluster randomised feasibility trial evaluating nutritional interventions in the treatment of malnutrition in care home adult residents. *Trials* 2015;16:433. doi:10.1186/s13063-015-0952-2.
33 Dickersin K, Manheimer E, Wieland S, Robinson KA, Lefebvre C, McDonald S. Development of the Cochrane Collaboration's CENTRAL Register of controlled clinical trials. *Eval Health Prof* 2002;25:38-64. doi:10.1177/0163278702025001004.
34 Leblanc A, Légaré F, Labrecque M, et al. Feasibility of a randomised trial of a continuing medical education program in shared decision-making on the use of antibiotics for acute respiratory infections in primary care: the DECISION+ pilot trial. *Implement Sci* 2011;6:5. doi:10.1186/1748-5908-6-5.

35 Hopewell S, Clarke M, Moher D, et al. CONSORT Group. CONSORT for reporting randomised trials in journal and conference abstracts. *Lancet* 2008;371:281-3. doi:10.1016/S0140-6736(07)61835-2.
36 Hopewell S, Clarke M, Moher D, et al. CONSORT Group. CONSORT for reporting randomized controlled trials in journal and conference abstracts: explanation and elaboration. *PLoS Med* 2008;5:e20. doi:10.1371/journal.pmed.0050020.
37 Heazell AE, Bernatavicius G, Roberts SA, et al. A randomised controlled trial comparing standard or intensive management of reduced fetal movements after 36 weeks gestation—a feasibility study. *BMC Pregnancy Childbirth* 2013;13:95. doi:10.1186/1471-2393-13-95.
38 World Medical Association declaration of Helsinki. Recommendations guiding physicians in biomedical research involving human subjects. *JAMA* 1997;277:925-6. doi:10.1001/jama.1997.03540350075038.
39 Hill KM, Walwyn RE, Camidge DC, et al. Lifestyle referral assessment in an acute cardiology setting: study protocol for a randomized controlled feasibility trial. *Trials* 2013;14:212. doi:10.1186/1745-6215-14-212.
40 Hill K, Walwyn R, Camidge D, et al. A randomized feasibility trial of a new lifestyle referral assessment versus usual assessment in an acute cardiology setting. *J Cardiovasc Nurs* 2015. doi:10.1097/JCN.0000000000000294.
41 Eldridge SM, Costelloe CE, Kahan BC, Lancaster GA, Kerry SM. How big should the pilot study for my cluster randomised trial be? *Stat Methods Med Res* 2016;25:1039-56. doi:10.1177/0962280215588242.
42 Clark WF, Sontrop JM, Huang S-H, et al. The chronic kidney disease Water Intake Trial (WIT): results from the pilot randomised controlled trial. *BMJ Open* 2013;3:e003666. doi:10.1136/bmjopen-2013-003666.
43 Buse GL, Bhandari M, Sancheti P, et al. Hip Fracture Accelerated Surgical Treatment and Care Track (HIP ATTACK) Investigators. Accelerated care versus standard care among patients with hip fracture: the HIP ATTACK pilot trial. *CMAJ* 2014;186:E52-60. doi:10.1503/cmaj.130901.
44 Busse M, Quinn L, Debono K, et al. Members of the COMMET-HD Management Group. A randomized feasibility study of a 12-week community-based exercise program for people with Huntington's disease. *J Neurol Phys Ther* 2013;37:149-58. doi:10.1097/NPT.0000000000000016.
45 Goldstein TR, Fersch-Podrat R, Axelson DA, et al. Early intervention for adolescents at high risk for the development of bipolar disorder: pilot study of Interpersonal and Social Rhythm Therapy (IPSRT). *Psychotherapy (Chic)* 2014;51:180-9. doi:10.1037/a0034396.
46 Schoultz M, Atherton I, Watson A. Mindfulness-based cognitive therapy for inflammatory bowel disease patients: findings from an exploratory pilot randomised controlled trial. *Trials* 2015;16:379. doi:10.1186/s13063-015-0909-5.
47 Hughes CW, Barnes S, Barnes C, Defina LF, Nakonezny P, Emslie GJ. Depressed Adolescents Treated with Exercise (DATE): A pilot randomized controlled trial to test feasibility and establish preliminary effect sizes. *Ment Health Phys Act* 2013;6:119-31. doi:10.1016/j.mhpa.2013.06.006.
48 Hoffmann TC, Glasziou PP, Boutron I, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *BMJ* 2014;348:g1687. doi:10.1136/bmj.g1687.
49 Boogerd EA, Noordam C, Kremer JA, Prins JB, Verhaak CM. Teaming up: feasibility of an online treatment environment for adolescents with type 1 diabetes. *Pediatr Diabetes* 2014;15:394-402. doi:10.1111/pedi.12103.
50 Begh RA, Aveyard P, Upton P, et al. Promoting smoking cessation in Pakistani and Bangladeshi men in the UK: pilot cluster randomised controlled trial of trained community outreach workers. *Trials* 2011;12:197. doi:10.1186/1745-6215-12-197.
51 National Institute for Health Research. Glossary | Pilot studies 2015 [cited 2015 30 June]. www.nets.nihr.ac.uk/glossary?result_1655_result_page=P.
52 Bugge C, Williams B, Hagen S, et al. A process for Decision-making after Pilot and feasibility Trials (ADePT): development following a feasibility study of a complex intervention for pelvic organ prolapse. *Trials* 2013;14:353. doi:10.1186/1745-6215-14-353.
53 Taylor SJ, Sohanpal R, Bremner SA, et al. Self-management support for moderate-to-severe chronic obstructive pulmonary disease: a pilot randomised controlled trial. *Br J Gen Pract* 2012;62:e687-95. doi:10.3399/bjgp12X656829.
54 Hamilton FL, Hornby J, Sheringham J, et al. DIgital Alcohol Management ON Demand (DIAMOND): feasibility randomised controlled trial of a web-based intervention to reduce alcohol consumption in people with hazardous and harmful use versus a face-to-face intervention: protocol. *Pilot Feasibility Stud* 2015;1:28doi:10.1186/s40814-015-0023-1.
55 Julious SA. Sample size of 12 per group rule of thumb for a pilot study. *Pharm Stat* 2005;4:287-91. doi:10.1002/pst.185.
56 Teare MD, Dimairo M, Shephard N, Hayman A, Whitehead A, Walters SJ. Sample size requirements to estimate key design parameters from external pilot randomised controlled trials: a simulation study. *Trials* 2014;15:264. doi:10.1186/1745-6215-15-264.
57 Browne RH. On the use of a pilot sample for sample size determination. *Stat Med* 1995;14:1933-40. doi:10.1002/sim.4780141709.
58 Whitehead AL, Julious SA, Cooper CL, Campbell MJ. Estimating the sample size for a pilot randomised trial to minimise the overall trial sample size for the external pilot and main trial for a continuous outcome variable. *Stat Methods Med Res* 2016;25:1057-73. doi:10.1177/0962280215588241.
59 Cocks K, Torgerson DJ. Sample size calculations for pilot randomized trials: a confidence interval approach. *J Clin Epidemiol* 2013;66:197-201. doi:10.1016/j.jclinepi.2012.09.002.
60 Warren FC, Stych K, Thorogood M, et al. Evaluation of different recruitment and randomisation methods in a trial of general practitioner-led interventions to increase physical activity: a randomised controlled feasibility study with factorial design. *Trials* 2014;15:134. doi:10.1186/1745-6215-15-134.
61 Moher D, Hopewell S, Schulz KF, et al. Consolidated Standards of Reporting Trials Group. CONSORT 2010 Explanation and Elaboration: Updated guidelines for reporting parallel group randomised trials. *J Clin Epidemiol* 2010;63:e1-37. doi:10.1016/j.jclinepi.2010.03.004.
62 Crawley E, Mills N, Beasant L, et al. The feasibility and acceptability of conducting a trial of specialist medical care and the Lightning Process in children with chronic fatigue syndrome: feasibility randomized controlled trial (SMILE study). *Trials* 2013;14:415. doi:10.1186/1745-6215-14-415.
63 Cook DJ, Rocker G, Meade M, et al. PROTECT Investigators Canadian Critical Care Trials Group. Prophylaxis of Thromboembolism in Critical Care (PROTECT) Trial: a pilot study. *J Crit Care* 2005;20:364-72. doi:10.1016/j.jcrc.2005.09.010.
64 Schunter M, Chu H, Hayes TL, et al. Randomized pilot trial of a synbiotic dietary supplement in chronic HIV-1 infection. *BMC Complement Altern Med* 2012;12:84. doi:10.1186/1472-6882-12-84.
65 Arnold DM, Heddle NM, Carruthers J, et al. A pilot randomized trial of adjuvant rituximab or placebo for nonsplenectomized patients with immune thrombocytopenia. *Blood* 2012;119:1356-62. doi:10.1182/blood-2011-08-374777.
66 Sackett DL. Commentary: Measuring the success of blinding in RCTs: don't, must, can't or needn't? *Int J Epidemiol* 2007;36:664-5. doi:10.1093/ije/dym088.
67 Forero M, Heikkila A, Paul JE, Cheng J, Thabane L. Lumbar transversus abdominis plane block: the role of local anesthetic volume and concentration—a pilot, prospective, randomized, controlled trial. *Pilot Feasibility Stud* 2015;1:10. doi:10.1186/s40814-015-0002-6.
68 Seebacher B, Kuisma R, Glynn A, Berger T. Rhythmic cued motor imagery and walking in people with multiple sclerosis: a randomised controlled feasibility study. *Pilot Feasibility Stud* 2015;1:25. doi:10.1186/s40814-015-0021-3.
69 Bruhn H, Bond CM, Elliott AM, et al. Pharmacist-led management of chronic pain in primary care: results from a randomised controlled exploratory trial. *BMJ Open* 2013;3:e002361. doi:10.1136/bmjopen-2012-002361.
70 Hopewell S, Hirst A, Collins GS, Mallett S, Yu LM, Altman DG. Reporting of participant flow diagrams in published reports of randomized trials. *Trials* 2011;12:253. doi:10.1186/1745-6215-12-253.
71 Van Spall HG, Toren A, Kiss A, Fowler RA. Eligibility criteria of randomized controlled trials published in high-impact general medical journals: a systematic sampling review. *JAMA* 2007;297:1233-40. doi:10.1001/jama.297.11.1233.
72 Burton C, Weller D, Marsden W, Worth A, Sharpe M. A primary care Symptoms Clinic for patients with medically unexplained symptoms: pilot randomised trial. *BMJ Open* 2012;2:e000513. doi:10.1136/bmjopen-2011-000513.
73 Egger M, Jüni P, Bartlett C. CONSORT Group (Consolidated Standards of Reporting of Trials). Value of flow diagrams in reports of randomized controlled trials. *JAMA* 2001;285:1996-9. doi:10.1001/jama.285.15.1996.
74 Vist GE, Hagen KB, Devereaux PJ, Bryant D, Kristoffersen DT, Oxman AD. Systematic review to determine whether participation in a trial influences outcome. *BMJ* 2005;330:1175. doi:10.1136/bmj.330.7501.1175.
75 Boysen G, Krarup L-H, Zeng X, et al. ExStroke Pilot Trial Group. ExStroke Pilot Trial of the effect of repeated instructions to improve physical activity after ischaemic stroke: a multinational randomised controlled clinical trial. *BMJ* 2009;339:b2810. doi:10.1136/bmj.b2810.
76 Frech SA, Dupont HL, Bourgeois AL, et al. Use of a patch containing heat-labile toxin from Escherichia coli against travellers' diarrhoea: a phase II, randomised, double-blind, placebo-controlled field trial. *Lancet* 2008;371:2019-25. doi:10.1016/S0140-6736(08)60839-9.
77 Olsen K, Howel D, Barber R, et al. Lessons from a pilot and feasibility randomised trial in depression (Blood pressure Rapid Intensive Lowering And Normal Treatment for Mood and cognition in persistent depression (BRILiANT mood study)). *Pilot Feasibility Stud* 2015;1:44. doi:10.1186/s40814-015-0042-y.

78 Ormiston JA, Serruys PW, Regar E, et al. A bioabsorbable everolimus-eluting coronary stent system for patients with single de-novo coronary artery lesions (ABSORB): a prospective open-label trial. *Lancet* 2008;371:899-907. doi:10.1016/S0140-6736(08)60415-8.
79 Thomas LH, Watkins CL, Sutton CJ, et al. ICONS Project Team and the ICONS Patient, Public and Carer Involvement Groups. Identifying continence options after stroke (ICONS): a cluster randomised controlled feasibility trial. *Trials* 2014;15:509. doi:10.1186/1745-6215-15-509.
80 Kapur N, Gunnell D, Hawton K, et al. Messages from Manchester: pilot randomised controlled trial following self-harm. *Br J Psychiatry* 2013;203:73-4. doi:10.1192/bjp.bp.113.126425.
81 Farrin A, Russell I, Torgerson D, Underwood M. UK BEAM Trial Team. Differential recruitment in a cluster randomized trial in primary care: the experience of the UK back pain, exercise, active management and manipulation (UK BEAM) feasibility study. *Clin Trials* 2005;2:119-24. doi:10.1191/1740774505cn073oa.
82 Zisk JL, Mackley A, Clearly G, Chang E, Christensen RD, Paul DA. Transfusing neonates based on platelet count vs. platelet mass: a randomized feasibility-pilot study. *Platelets* 2014;25:513-6. doi:10.3109/09537104.2013.843072.
83 Pai M, Lloyd NS, Cheng J, et al. Strategies to enhance venous thromboprophylaxis in hospitalized medical patients (SENTRY): a pilot cluster randomized trial. *Implement Sci* 2013;8:1. doi:10.1186/1748-5908-8-1.
84 Jackson L, Cohen J, Sully B, Julious S. NOURISH, Nutritional OUtcomes from a Randomised Investigation of Intradialytic oral nutritional Supplements in patients receiving Haemodialysis: a pilot randomised controlled trial. *Pilot Feasibility Stud* 2015;1:11. doi:10.1186/s40814-015-0007-1.
85 Kennedy CC, Thabane L, Ioannidis G, Adachi JD, Papaioannou A. ViDOS Investigators. Implementing a knowledge translation intervention in long-term care: feasibility results from the Vitamin D and Osteoporosis Study (ViDOS). *J Am Med Dir Assoc* 2014;15:943-5. doi:10.1016/j.jamda.2014.05.007.
86 Alers NO, Jenkin G, Miller SL, Wallace EM. Antenatal melatonin as an antioxidant in human pregnancies complicated by fetal growth restriction--a phase I pilot clinical trial: study protocol. *BMJ Open* 2013;3:e004141. doi:10.1136/bmjopen-2013-004141.
87 Dyer C. UK clinical trials must be registered to win ethics committee approval. *BMJ* 2013;347:f5614. doi:10.1136/bmj.f5614.
88 Kmietowicz Z. NHS research authority links approval of trials to registration and publication of results. *BMJ* 2013;346:f3119. doi:10.1136/bmj.f3119.
89 World Health Organization. WHO | Welcome to the WHO ICTRP 2015 [cited 2015 30 June]. www.who.int/ictrp/en/.
90 International Committee of Medical Journal Editors. ICMJE | Clinical Trials Registration 2015 [cited 2015 30 June]. www.icmje.org/about-icmje/faqs/clinical-trials-registration/.
91 Hecksteden A, Grütters T, Meyer T. Associations between acute and chronic effects of exercise on indicators of metabolic health: a pilot training trial. *PLoS One* 2013;8:e81181. doi:10.1371/journal.pone.0081181.
92 Yu X, Stewart SM, Chui JP, Ho JL, Li AC, Lam TH. A pilot randomized controlled trial to decrease adaptation difficulties in chinese new immigrants to Hong Kong. *Behav Ther* 2014;45:137-52. doi:10.1016/j.beth.2013.10.003.
93 Items SP. Recommendation for Interventional Trials. SPIRIT | Welcome to the SPIRIT Statement website 2013 [cited 2015 30 June]. www.spirit-statement.org/.
94 Hip Fracture Accelerated Surgical Treatment and Care Track (HIP ATTACK) Investigators. Accelerated care versus standard care among patients with hip fracture: the HIP ATTACK pilot trial. *CMAJ* 2014;186:E52-60. doi:10.1503/cmaj.130901.
95 Porserud A, Sherif A, Tollbäck A. The effects of a physical exercise programme after radical cystectomy for urinary bladder cancer. A pilot randomized controlled trial. *Clin Rehabil* 2014;28:451-9. doi:10.1177/0269215513506230.
96 Wilson DT, Walwyn RE, Brown J, Farrin AJ, Brown SR. Statistical challenges in assessing potential efficacy of complex interventions in pilot or feasibility studies. *Stat Methods Med Res* 2016;25:997-1009. doi:10.1177/0962280215589507.
97 Turner L, Shamseer L, Altman DG, Schulz KF, Moher D. Does use of the CONSORT Statement impact the completeness of reporting of randomised controlled trials published in medical journals? A Cochrane review. *Syst Rev* 2012;1:60. doi:10.1186/2046-4053-1-60.
98 Penelope 2016 [18 March 2016]. www.peneloperesearch.com.

Supplementary file: CONSORT checklist of information to include when reporting a pilot trial

BMC Public Health

CORRESPONDENCE Open Access

How to develop a theory-driven evaluation design? Lessons learned from an adolescent sexual and reproductive health programme in West Africa

Sara B Van Belle*, Bruno Marchal, Dominique Dubourg, Guy Kegels

Abstract

Background: This paper presents the development of a study design built on the principles of theory-driven evaluation. The theory-driven evaluation approach was used to evaluate an adolescent sexual and reproductive health intervention in Mali, Burkina Faso and Cameroon to improve continuity of care through the creation of networks of social and health care providers.

Methods/design: Based on our experience and the existing literature, we developed a six-step framework for the design of theory-driven evaluations, which we applied in the ex-post evaluation of the networking component of the intervention. The protocol was drafted with the input of the intervention designer. The programme theory, the central element of theory-driven evaluation, was constructed on the basis of semi-structured interviews with designers, implementers and beneficiaries and an analysis of the intervention's logical framework.

Discussion: The six-step framework proved useful as it allowed for a systematic development of the protocol. We describe the challenges at each step. We found that there is little practical guidance in the existing literature, and also a mix up of terminology of theory-driven evaluation approaches. There is a need for empirical methodological development in order to refine the tools to be used in theory driven evaluation. We conclude that ex-post evaluations of programmes can be based on such an approach if the required information on context and mechanisms is collected during the programme.

Background

Theory-driven evaluation (TDE) was invented to provide an answer to problems of evaluation approaches that are limited to before-after and input-output designs traditionally used in programme evaluation [1,2]. This was a reaction to the position of Campbell & Stanley [3], who stated that internal validity is the most essential issue in research, and Cronbach's position that evaluation cannot serve policymaking if its external validity is not guaranteed [4]. Chen and Rossi aimed at providing a perspective on evaluation that ensures both types of validity. These authors hold that for any intervention, a programme theory that explains how the planners expect the intervention to work can be described. The programme theory is the often implicit set of assumptions that steers the choice and design of an intervention. Making these assumptions explicit allows to understand what is being implemented and why - it opens up the so-called black box between intervention and outcome. Therefore, the programme theory represents a hypothesis that can be tested and further refined.

Chen distinguishes the normative from the causal part of the programme theory [1]. The normative theory or *action model* contains the rationale and justification of the programme [5]. It is what programme designers have in mind as assumptions and objectives when designing the programme. In many programmes, these assumptions are not stated explicitly; it is simply assumed that all programme partners share them. Evaluation of the action model describes how exactly the planned intervention has been implemented and allows

* Correspondence: svanbelle@itg.be
Institute of Tropical Medicine, Nationalestraat 155, B-2000 Antwerp, Belgium

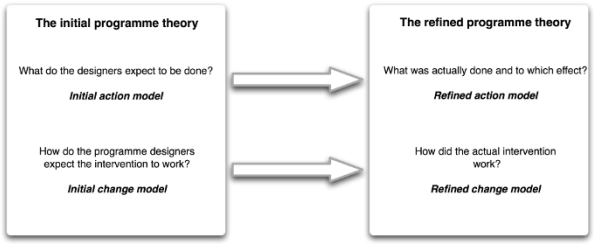


Figure 1 The components of the initial and the refined programme theory.

to check whether an unsuccessful intervention is due to implementation failure or programme design failure. Evaluation of the causal theory or *change model* examines the causal processes and the intervening contextual variables that produce change [5] (see figure 1). In theory-driven evaluation, the results of the evaluation are formulated as an improved programme theory and as such incorporated into the existing body of theoretical and programme knowledge.

Since the 1990s, new developments in the field of theory-driven evaluation include Theory of Change and realist evaluation. The Theory of Change (ToC) approach was developed by the Roundtable on Community Change (Aspen Institute, USA) to evaluate complex community-based change interventions [6]. Mostly applied to programme evaluation, it seeks to establish the links between intervention, context and outcome [7-9] through development and testing of logic models [10].

Realist Evaluation (RE), developed by Pawson & Tilley [11], argues that in order to be useful for decision makers, evaluations need to indicate 'what works in which conditions for whom', rather than merely answering the question 'does it work?'. Realist evaluation aims at identifying the underlying generative mechanisms of the intervention, the "pivot around which RE evolves", [12] - and the influence of context upon the outcomes. It has its philosophical roots in critical realism [13,14]. In this paper, we adhere to the theory-driven evaluation terminology of Chen [15] for reasons of simplicity, acknowledging the wide range of other terms used in Theory of Change and Realist Evaluation.

Theory-driven evaluation somehow disappeared from the radar during the 1990s, emerging again at the European Conference of Evaluation in 2002 [16]. Methodological developments had continued, however, in the field of programme evaluation by authors like Chen [5] and Donaldson [17]. In parallel, ToC and Realist Evaluation were increasingly applied in the evaluation of social care, youth and education policies and programmes [8,18-23].

In health care, there is limited documentation regarding the practical application of research and evaluation designs based on theory-driven evaluation principles. In the domain of health promotion, there are some studies using the ToC approach [21] or Realist Evaluation [24,25]. In the field of health policy and management, papers include [26,27] and [28]. In the domain of medical education, we found two papers ([12,29]). There are even fewer publications focusing on the practical application in public health in low and middle-income countries [21]. These include some research studies in the domain of health service organisation and public health ([30,31]).

This scarcity of theory-driven enquiry in health may be due to various reasons: carrying out a full-blown theory-driven evaluation is resource- and time intensive [2]. Furthermore, guidance on how to apply the principles of theory-driven evaluation in the domain of health systems research is scarce. Indeed, few of the abovementioned papers offer structured approaches to practically carrying out such evaluations or research.

The objective of the evaluation on which we report was not only to assess the intervention, but also to systematically develop a framework for the design of theory-driven evaluations.

We first describe briefly the programme that was evaluated and then present how we developed a 6-step framework for the design of a theory-driven evaluation protocol. For each step, we describe how we applied it during the evaluation. We end by discussing the main challenges we faced, framing our experience in the existing literature.

Methods/design

We applied the principles of theory-driven evaluation in an ex-post evaluation of one of the programme strategies of the PASSAGE programme, Projet d'Approche Solidaire en Santé Génésique. PASSAGE is a EU funded, three-year intervention aiming at improving the continuity of care in reproductive health in an urban setting in Mopti (Mali), Maroua, (Cameroon) and two districts of Ouagadougou (Burkina Faso), which ran from 2006 to 2009.

The object of the evaluation was the creation of networks between public and private health and social service providers in adolescent sexual and reproductive health. These networks aimed at improving the integration of services and the continuity of care for adolescents.

Based on our experience and existing literature [5,15,17], we developed a six-step framework for the design of theory-driven evaluations in the field of health systems:

- Step 1: Assessing the scope of the evaluation and the appropriateness of TDE
- Step 2: Critical reconstruction of the initial programme theory
- Step 3: Choice of data collection methods & development of tools
- Step 4: Assessing the initial action model: Evaluating relevance of programme design and degree of implementation
- Step 5: Assessing the initial causal model: Establishing the causal mechanisms and contextual factors, and their interactions
- Step 6: Translating findings into the refined programme theory

Step 1: Assessing the scope of the evaluation: Is TDE needed in order to learn?

Theory-driven evaluation can be quite resource- and time intensive: its scope extends beyond an efficacy/outcome evaluation to include the assessment of the underlying programme theory [32]. Also the need to deconstruct the influence of the context on the intervention and the outcomes requires time [33]. It is therefore important to assess the scope of the evaluation to decide whether a TDE approach is needed. A number of authors have indicated the usefulness of TDE in evaluation of interventions that have attributes of complexity [7,18,34,35]. We argue here that TDE can be used to good effect in case of research or evaluation of an intervention in a complex setting and in case of a new type of intervention, for which the understanding of the causal mechanisms needs to be established.

In practice, the need for a TDE approach for the evaluation of the networking component of PASSAGE was jointly assessed with the commissioner of the evaluation. We found that the evaluation of the networking strategy fulfilled the above indications: it is an intervention in a complex setting where social interaction needs to be mobilized for the intervention to succeed. In order to improve continuity of care for adolescent sexual and reproductive health, PASSAGE intended to create or strengthen linkages between professional and non-professional service providers of different sectors: public and private, medical and social. The creation of networks between these different communities of providers intervening at different levels inside and outside of the health system requires the initiation and maintenance of a social dynamic between them. It could also be argued that the networking component was innovative, and thus requiring in-depth investigation. The creation or promotion of networking is a tested intervention in the field of development (e.g. the creation of national NGO platforms in Sub Sahara Africa) and in public health (e.g. the creation of networks of HIV/AIDS civil society organisations). However, networking has seldom been applied to stimulate (promote) integrated care provision in the domain of reproductive health.

During this step, it was also decided to mainly focus this evaluation on the processes through which the intervention worked (or not). The specific objective was to evaluate to what extent strategies implemented to strengthen networks between actors involved in adolescent sexual and reproductive health (ASRH) contributed to:

- the creation of a common vision on an integrated approach towards ASRH service delivery among the involved service providers
- strengthening the capacities of associations involved in the network and improving their functioning
- an improved integration of services and better continuity of care
- better collaboration between the Regional Directorate of Health, one of the programme's implementing partners, and the networks created or revitalised by the programme.

Step 2: Critical reconstruction of the initial programme theory

A second step in a TDE evaluation is to make the initial programme theory (PT) explicit, the - often implicit - assumptions of the actors involved in the design and subsequent implementation of the intervention. They include the programme designers and implementation teams in each setting, partners in implementation and the target group, in this case the adolescents. Describing the initial PT explicitly aims at understanding the actors' interpretations of how the intervention is linked to the outcomes through eliciting their assumptions regarding the underlying mechanisms.

Lipsey & Pollard identify different mechanisms to make this PT explicit [36]. First, much relevant information can be gained from the designers and implementers. In this case, the researchers unearth the models that the actors are using to describe and understand the intervention through individual interviews or group discussions. Cole stresses the need to involve the stakeholders and implementers of the interventions during the stage of programme theory development, as one seeks to describe what these actors think compared with what the designers thought [37]. The discrepancy between these views may then be explored as a source of non-implementation [17].

A second source of information for constructing the initial programme theory is relevant theories and current knowledge, such as findings from evaluations of

similar interventions. In some cases, the problem situation, the intervention or the policy has been thoroughly researched. The results of these studies can contribute to the formulation of the PT. In other cases, appropriate concepts from disciplines such as cognitive psychology, sociology, etc. can be used [36].

A third approach consists of exploratory on-site research during the various phases of the programme based on observation and inquiry. In all three cases, the feedback of the emerging programme theory to the actors involved is critical, since this allows refining it [36]. In practice, the three approaches are used in combination (see for instance [28]).

When programmes are evaluated, a natural starting point is the logic model presented by the logical framework. In practice, however, the logical framework often offers little information on mechanisms of change. Also, they are usually developed before the start of the programme without much consultation of the implementers or beneficiaries. This lack of useful information often persists, since once the programme starts, there is frequently too little time to build a shared understanding of the logical framework. In such case, the actors typically rely on their own interpretation of the logical framework and this provides the main guidance for implementation [21,34]. If this is the case, one might find that several rival programme theories co-exist and evaluators will need to explore these different interpretations. At the least, they should establish to which degree the initial programme theory was shared by the main actors.

In the evaluation of PASSAGE, we started to draft the programme theory by reviewing the main programme documents, such as the description of the intervention in the programme proposal and the logical framework. We then interviewed the programme coordinator, who also was the main initiator and designer. We explored the literature to frame the programme designers' assumptions against the existing theory.

To structure the initial PT, we used the following elements: the planned intervention and its elements, the planned outputs and outcomes, the context factors assumed to be needed and the processes of change. Table 1 presents the initial PT that was the result of the above process. In a second stage, the programme theory as perceived by each country programme implementation team and by the implementing partners was generated. Divergent interpretations and adaptations to the context, indeed, need to be identified as they may pull the programme's implementation in different directions. To this end, the teams and partners were interviewed. In a third phase, we interviewed adolescents in each site.

Due to the nature of the intervention, e.g. the large number of actors and associations involved, and the limited time spent at the start of the programme on building a joint understanding of the logical framework, we expected that divergent perspectives on the programme theory would emerge. During the design phase, we decided therefore to describe any such rival PT and compare them in the analysis phase of the evaluation. In practice, we found that the PT of the country programme teams did not significantly differ from the overall PASSAGE PT described in Table 1, but as we will see below, the activities that were actually conducted were different across the sites.

Step 3: Choice of data collection methods & development of tools

Theory-driven evaluation is essentially method-neutral. Both quantitative and qualitative data collection methods can be used. The choice of data collection methods and the actual data collection process is steered by the aim of the study, its scope and the degree of development of the programme theory: the aim is to collect data to confirm or falsify its different elements and linkages [11].

In this case, we chose for the case study as the overall design, a natural choice for the evaluation of a programme component (in this case 'networking') in which social dynamics are assumed to be important. The case study design allows for exploring a "phenomenon within its real-life context, especially when the boundaries

Table 1 The initial programme theory of PASSAGE

The initial action model (What did programme designers plan to do and expect to attain?)	Bringing together the various actors involved in reproductive health for adolescents in a network increases the access and the utilisation of appropriate social and health services by adolescents and contributes to improving their sexual and reproductive health status.
The Initial change model (How was the programme supposed to work based on the programme designers' assumptions?)	The network(ing) contributes to: (1) better knowledge of partners with different backgrounds and their specificities; (2) a growing awareness of a shared vision among partners on adolescent sexual and reproductive health; Knowing each other and each others' specificities and a growing awareness of a shared vision would lead to cooperation and the creation of synergies rather than competition. This in turn would lead to improved ASRH outcomes.

between phenomenon and context are not clearly evident" [38].

Given the main focus of the evaluation on the causative theory, mostly qualitative methods were to be used: both the identification of the key elements of the programme theory as the exploration of underlying mechanisms required interviews and focus group discussions, besides the analysis of progress reports and logbooks.

In practice, this step coincided somehow with step 2, as at that stage, tools were needed for programme document review, the literature review and the interviews with the programme designers. Semi structured topic guides were drafted for these interviews. We discuss the specific data collection issues for step 4 and 5 below.

Step 4: From initial action model to refined action models: evaluating the relevance of programme design and strength of implementation in the three settings

Once the initial programme theory has been described, the data collection tools designed and the data collected, the programme theory can be used in the next step: the actual assessment of the different dimensions of the programme in function of the actual research questions.

First, the evaluators focus on the action model, describing the programme design on the one hand and its actual implementation on the other. This step assesses the congruency between the planned and the actual intervention and looks at potential issues concerning implementation. It allows distinguishing theory failure from implementation failure [5].

In the case of PASSAGE, we designed the protocol to provide answers to the following questions in each of the study sites:

(1) What was the actual intervention implemented as compared to the planned intervention?

(2) How was the intervention implemented?

(3) What were the results of the intervention?

To this end, the data collection was carried out in the three study settings. In preparation of the fieldwork, a primer in theory-driven evaluation was designed and used for training of the local research teams. Interview guides were drafted, fine-tuned and tested in each field site prior to the actual interviews. A team consisting of 2 evaluators carried out the fieldwork during a 2-week period at each study site.

At the start of the fieldwork, the country-level programme theory was formulated on the basis of interviews with the country programme team members. In a second step, in-depth interviews were carried out with purposively selected key informants in order to obtain information on the actual implementation of the programme, the mechanisms and context (see Table 2). These included representatives of the local authorities and the district and regional health authorities, staff of public and private health facilities, staff and volunteers in youth centres. We also interviewed members of community-based organisations and NGOs involved in adolescent sexual and reproductive health, peer educators and volunteers of school youth groups and neighbourhood youth groups, community and religious leaders. In a third step, focus groups discussions were carried out, the participants of which were divided by sex in separate groups. The age of the participants was between 15 and 24 years. Each group was selected to contain adolescents of different substrata: adolescents from various neighbourhoods, adolescents going to school and being out of school, adolescents following comprehensive education and technical (professional) education, adolescents from private, faith based schools and from public schools.

Additional information on programme implementation was obtained by reviewing the progress reports and the logbook kept by the programme coordinator.

Step 5: From initial change model to refined change models: establishing the causal mechanisms and contextual factors in the three settings

Theory-driven evaluation would not provide an added value compared to result-based (outcome/impact) evaluations if the change model would be left unchecked. This step traces the mechanisms that link the actual intervention to the actual outcomes. By mechanism, we understand the causal pathway that is made up by the interplay between intervention, actors and contextual conditions. This interplay may consist of both linear relations and feedback loops that ultimately lead to change.

The evaluation of the change model answers three questions: (1) What kind of relationships exist between actual intervention and outcome?; (2) Which intervening factors could be mediating the effect of the intervention on the outcome variables? and (3) Under what contextual conditions will the causal relationship be facilitated or inhibited? [5].

The actual intervention

In the case of the PASSAGE evaluation, we proceeded by first describing the networking component of the programme as it was actually implemented on the ground in each site. We found that the actual networking component differed across the sites (Table 3). Also the speed of their development varied. In Mopti (Mali), it took some time to warm NGOs to the idea of a reproductive health network and during the evaluation, network members were still in the process of exploring the possibilities.

Table 2 Overview of in-depth interviews and focus group discussions

	Mali	Burkina Faso	Cameroun
In-depth interviews	25	24	25
Focus group discussions	1 with 8 male adolescents	1 with 10 male adolescents	1 with 9 male adolescents
	1 with 8 female adolescents	1 with 10 female adolescents	1 with 9 female adolescents

The results

In a second step, we assessed the results of the intervention.

- In general, we found that the intervention in the three settings resulted in closer collaboration between partners who before the programme were only loosely connected.
- We found that the exchange between different NGOs during network meetings resulted in several activities that brought together NGOs and technical services or formal health structures. In Burkina Faso, for instance, the networks launched by the project have resulted in a better ad hoc referral between ASRH curative and preventive services. Meetings were organized that bring together all actors of both public and private non-for-profit sector. At the Cameroon intervention site, all youth volunteers active in reproductive health and working in schools were brought together, creating linkage and exchange between adolescents of different denominations and backgrounds.
- Our data indicates that the networking strategy led to increased organisational learning through exchange of information, expertise and material resources.
- It also led to an analysis of the offer of care and some remedial action. The health professionals became aware of lacunae in the provision of and access to ASRH services and that actions were taken that improved the continuity of care for adolescents.
- We also found that the DRS, who according to the plan was to take up the coordinator role, did support the networking activities in all three settings but did not fully take up the role of coordinator.

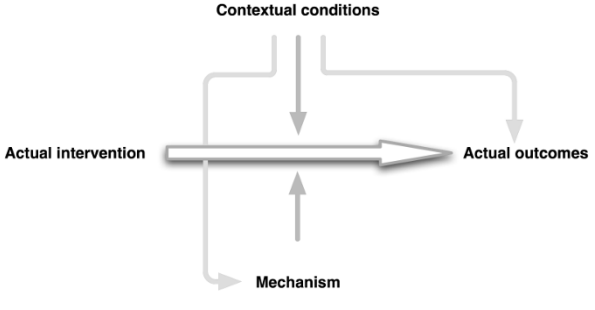


Figure 2 The change model.

Mechanisms

In a third step, we sought clues and information for these mechanisms during the in-depth interviews and observations. To do so, we included questions covering the following themes: the process of networking (the process of setting up networks or revitalisation of existing ones, network members and connections, activities organised by the networks, etc.); results of activities conducted by the network (sharing of knowledge, dialogue, improved coordination under the aegis of the regional health authorities, etc.), appropriateness of the networking strategy to the site context, and the sustainability of the networks.

We found that important factors were: (a) perceived individual and organizational opportunities and (b) an individual or organizational awareness of the lacunae in ASRH service delivery leading to a commitment to improve ASRH services. Individuals and organizations want to participate actively in a network when they perceive that this is of added value to their functioning. Network actors joined a network because it enhances their organizational visibility, to liaise and learn from other resource persons and organizations in the field (as most organizations are not specialized in ASRH and recognize that they are in need of additional expertise), to have access to information and training and, last but not least, to have access to additional funding opportunities.

Context

Fourth, we set out to describe the influence of the context. The literature shows that contextual conditions that facilitate or inhibit processes of change entail institutional arrangements, stakeholders' and target groups' attitudes and behaviours, and geographical and sociocultural factors, either at meso- or macro level. During the analysis, two conditions emerged from the data. These were related to the networking process and to the relationship between networking as an intervention and the outcome. An example of the latter is the urban setting of the project, which facilitates communication between network members. We found that the following

Table 3 The networking activities in the 3 sites

Mali	The project team decided to strengthen the functioning of an existing NGO network that grouped HIV/AIDS NGOs of the region. This network was in a fragile state due to lack of leadership. The team decided to expand its membership to NGOs working in sexual and reproductive health.
Burkina Faso	Networking efforts were focused on the improvement of access to ASRH services through filling in gaps in the referral chain. Two networks were launched: REPERE (*Réseau des Personnes Référentes*) and RESCOPE (*Réseau des Structures Communautaires pour la promotion de la Paire-Education*). REPERE brings together individuals, working in both public health structures and private non-for-profit associations, who volunteer to act as an entry point for information for adolescents in need of ASRH services. Volunteers can be contacted by adolescents when in need. RESCOPE and REPERE work in tandem: peer educators of different youth associations provide information themselves or could refer adolescents in need of youth friendly service providers.
Cameroon	Three networks were launched: one bringing together peer educators of existing school clubs that were working on ASRH and HIV/AIDS prevention, one resource persons network, and a network of NGOs/CBOs working on HIV/AIDS prevention.

contextual conditions are related to the networking process itself:

- The competition context determines the degree of the net benefits to networking for the actors concerned. In a highly competitive environment, where NGOs have to compete for scarce resources, it might well be that networking, and particularly the sharing of information with other NGOs in the same field, might be perceived as detrimental to the organization.
- The commitment of the Regional Directorate of Health to be part of the network, to coordinate (stimulate) it and to oversee the private-not-for-profit sector not only depends on the benefits of this role for itself. It requires resources to do so, and we found that in all three settings, the DRS currently lacks the necessary financial resources, both financial, human resources and time, to take up this role. Furthermore, given the resource poor context, taking on a stewardship role might prove not to be beneficial as this could have negative financial implications. The private-non-for-profit sector could ask for financial support for activities that are of mutual benefit.
- For the DRS to take up the coordination role in a non-hierarchical structure such as the PASSAGE networks, it has to be accepted by the non-for-profit sector as the steward in adolescent sexual and reproductive health. In the Burkina Faso setting, private non-for-profit actors saw the benefit of working alongside the DRS for medical supervision and technical assistance. This was not the case in the other settings.

We summarised the resulting analysis of this step in a diagram of the causal pathways, which was validated through discussion with the programme partners during the fieldwork and analysis phase.

Step 6: Generalization to the level of a refined programme theory

Theoretically, TDE yields results that have a higher external validity, because it ends with a refined programme theory that explains under which conditions and how the results were obtained. However, the literature does not provide us with much practical lead on how to generalize from particular evaluation findings. This is partly because of the non-linear, creative nature of theory constructions, where one goes back and forth between intuition and data, and between induction and deduction [39], a process that is hard to formalise.

In the case of evaluations, the refined PT should ideally make sense to the users of the evaluations and meet the purpose of the evaluation as defined by its commissioners. Furthermore, it needs to be able to serve as the starting point of evaluations of similar interventions, thus adding to an ever-increasing knowledge base regarding a particular intervention [39,40]. To this end, it should be formulated so as to explain not only whether the intervention works, but also how, for whom and in which context. In the case of PASSAGE, we ended by formulating the refined programme theory in a narrative form (Table 4).

Discussion

In this paper, we identified conditions that can be used to decide whether a theory-driven evaluation would be indicated. We discussed how the protocol was constructed around 6 steps that systematically apply the principles of theory-driven evaluation to an ex- post evaluation, presented the challenges and gave examples of the findings that emerged from the actual evaluation at each step.

During the design and implementation phase, we were confronted with several challenges. First, we faced the challenge of the variable and, at times, too vague terminology used by theory-driven evaluation experts and methodologists. Each major school develops its own terminology (see for instance 'middle range theory' [11], theory of change [6] or programme theory [15]; or normative and causative theory [15] versus action and causal model [5]. In many papers, the different approaches of theory of change, theory-driven evaluation and realist evaluation are somehow mixed up and terms of different schools are used interchangeably (see for instance [18,19,41].

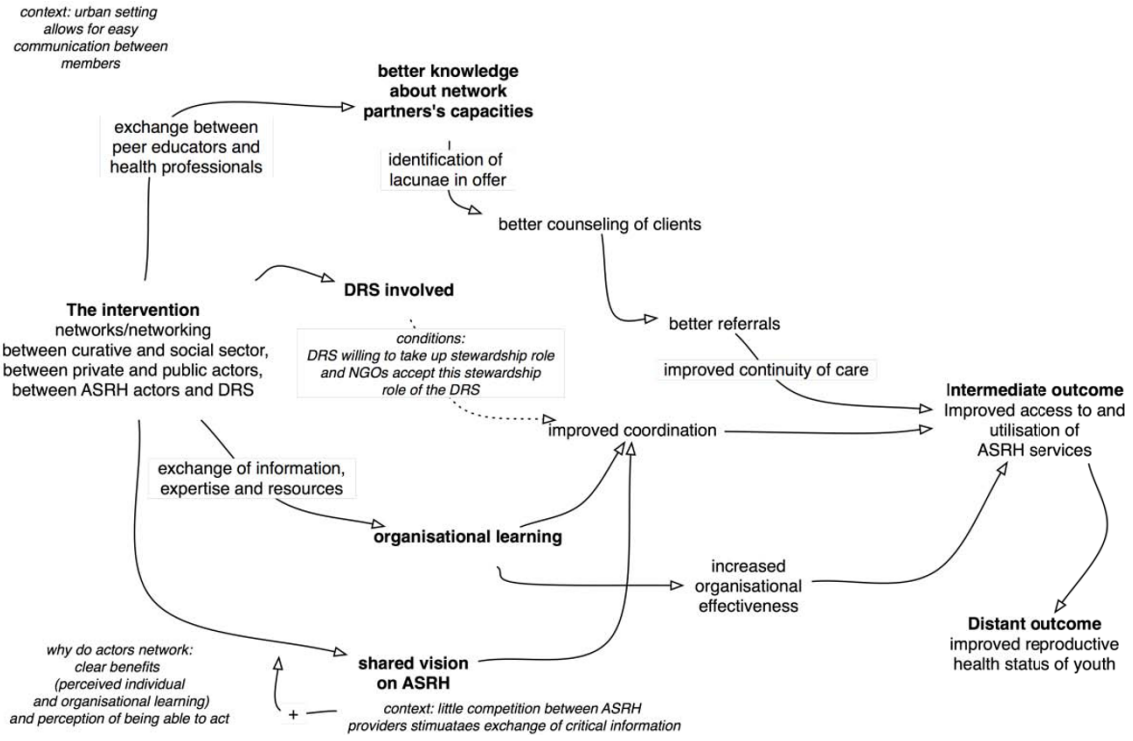


Figure 3 The causal web.

The issue of identifying 'rival' programme theories provides a good example of the limited published guidance. Rival PT are the result of actors' different viewpoints and positions vis-à-vis the intervention (for instance: initiator and designer versus implementer versus adolescents; the perspective from the South versus from the North). It is therefore important to identify whether any rival PT were held and how they influenced the programme. During the design phase, we realised that the heads of the country teams could have other interpretations of the goals and strategy of PASSAGE on the basis of their different professional backgrounds and experiences or personal preferences. We found some guidance in the literature: if different actors are gathered to discuss the programme theory at the programme start-up phase, the role of the evaluator will be one of negotiator between groups in an - in essence - political process [21,42]. If the evaluator is involved in the building of the M&E system at the beginning of the programme, clear responsibilities between the programme coordinator and the evaluator need to be delineated to avoid a blurring of roles between them [21].

Table 4 The refined PT

The refined programme theory of Passage	Bringing together the various actors involved in reproductive health for adolescents in a network can increase the access and the utilisation of appropriate social and health services by the adolescents and contributes to improving the reproductive health status of the adolescents if (1) it succeeds to bring together actors that cover the whole range of services required by adolescents, (2) creates a shared vision and (3) leads to integration of all ASRH services. Active networking contributes to: (1) a shared awareness that the current services are ineffectual because of gaps and redundancies in the provision (2) better knowledge of partners with different backgrounds and thus to better informing adolescents and to more effective referrals, which in turn contributes to better continuity of care (3) a shared vision among partners on ASRH, which contributes to better coordination and integration of services (3) organisational learning, which enhances coordination and quality of care and services. The underlying processes include increasing *linking* social capital and organisational social capital. The latter strengthens the relations between organisations, the former stresses the relations between organisations and public authorities. Partners need to perceive a win-win situation to continue to be active members and to experience a feeling of ownership. Existing networks can be mobilised to take on new tasks, inactive networks can be revitalised (but this requires more time and inputs), or completely new networks can be set up (the longest route).

As mentioned above, in PASSAGE, we decided to maintain any such rival theory as an alternative hypothesis to be tested during the analysis.

Other challenges relate to the application of the TDE approach to ex-post programme evaluations. In essence, routine M&E systems of programmes do not monitor the contextual conditions that may be important, nor do they provide information that could allow identifying the underlying mechanisms. Combined with the issue of recall bias, this presents major challenges. One could argue that TDE could still be applied if during the evaluation, the change processes are explored in a joint reflection process where all actors join in, for instance during an end-of-programme closure workshop. We would tend to believe that such discussions would yield interesting information but not allow for a robust evaluation. It could thus be argued that ex-post evaluations of Log Frame based programmes are not possible, or at least that a complete application in its full scope is not feasible. Only if appropriate monitoring systems are built in the programme can information to identify mechanisms and contextual conditions be available at the end of a project.

Finally, we faced some more general challenges. First, there is the issue of the role and the skills of the evaluators. Development intervention evaluators are commonly driven towards establishing the outcomes of the programme and focus on changes within the target group of the intervention. Theory-driven evaluation requires additional training or thorough briefings to modify the evaluator's point of view from an exclusively results-driven focus (i.e. as needed in effectiveness evaluations) to a process-oriented focus that is needed for theory-driven evaluation. We found that theory-driven evaluation teams ideally have broad competencies, experience and expertise that allow for the identification of mechanisms of change and of the relevant contextual factors.

Second, it is often argued that TDE is time consuming [2]. In practice, we found that a TDE approach should not necessarily take more time than regular evaluations of similar multi-country programmes. In the case of PASSAGE, the preparation of the evaluation by the political scientist took about 2 weeks time, including the design of the protocol and the primer on TDE used in the training of the anthropologists. The fieldwork took the evaluation team consisting of one political scientist and one anthropologist 2 weeks per site. The analysis was based on site reports written by the anthropologists (2 weeks per site) and the comparative analysis took 4 weeks, including the draft of the final report.

A third general challenge is the issue of complexity. One major setback and perhaps also a reason why there is currently not an abundance of theory-driven evaluations of public health interventions, is the challenge that 'complexity' presents to the causal attribution. Whereas we argued above that theory-driven evaluation designs are appropriate for complexity, the very complex nature of the programmes stands in the way of an easy assessment of effectiveness and of underlying mechanisms: the outcomes of complex interventions can be attributed to a number of determinants, only some of which are influenced by the intervention. It would however be a mistake to adopt the standards of strength of evidence from the biomedical world in assessing the value of findings of evaluations of complexity. As Pawson & Tilley argue convincingly, in such cases, theory-driven evaluation will aim at offering plausible explanations, not probabilistic statements [11].

Conclusions

To conclude, the theory-driven evaluation approach holds much promise for relevant learning from public health interventions and programmes, but there still is a need for methodological development for practical use. Ex-post evaluations of programmes can be based on such an approach if the required information on context and mechanisms is collected during the programme.

TDE inevitably requires an element of practitioner "craft", involving judgment and creativity based on broad theoretical induction and background, and experience. Ways to reduce the danger of arbitrariness, unwarranted subjectivity and superficiality include (1) introducing the theory-driven perspective from the start of the programme, and (2) documented critical exchanges among TDE practitioners on how they deal effectively with vagueness and conceptual ambiguity.

Authors' contributions

SVB contributed to the design of the evaluation protocol, the actual evaluation, the analysis of findings and the writing of the manuscript. BM contributed to the protocol development and the writing of the manuscript, DD participated in the protocol development and manuscript revision, and GK in manuscript writing and revision. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 4 August 2010 Accepted: 30 November 2010
Published: 30 November 2010

References

1. Chen H-T: **The conceptual framework of the theory-driven perspective.** *Evaluation and Program Planning* 1989, **12**:391-396.
2. Chen H-T, Rossi P: **Issues in the theory-driven perspective.** *Evaluation and Program Planning* 1989, **12(4)**:299-306.
3. Campbell D, Stanley J: *Experimental And Quasi-experimental Designs For Research* Skokie, Illinois: Rand-McNally; 1963.
4. Cronbach L: **Associates Toward reform or program evaluation: aims, methods and institutional arrangements.** San Francisco: Jossey-Bass; 1980.
5. Chen H-T: *Practical program evaluation* Thousand Oaks: SAGE Publications; 2005.

6. Fulbright-Anderson A, Kubisch A, Connell J: *New approaches to evaluating community initiatives* Washington, DC: Aspen Institute; 1998.
7. Barnes M, Matka E, Sullivan H: **Evidence, understanding and complexity. Evaluation in non-linear systems.** *Evaluation* 2003, **9(3)**:265-284.
8. Mason P, Barnes M: **Constructing theories of change.** *Evaluation* 2007, **13(2)**:151-170.
9. Weiss C: **Nothing as practical as good theory: exploring theory-based evaluation for comprehensive community initiatives.** In *New approaches to evaluating comprehensive community initiatives: concepts, methods and contexts.* Edited by: Connell J, Kubisch A, Schorr L, Weiss C. Washington, DC: The Aspen Institute; 1995.
10. Douglas FC, Gray DA, van Teijlingen ER: **Using a realist approach to evaluate smoking cessation interventions targeting pregnant women and young people.** *BMC Health Serv Res* 2010, **10**:49.
11. Pawson R, Tilley N: *Realistic Evaluation* London: Sage; 1997.
12. Ogrinc G, Batalden P: **Realist evaluation as a framework for the assessment of teaching about the improvement of care.** *J Nurs Educ* 2009, **48(12)**:661-667.
13. Connelly J: **A realistic theory of health sector management. The case for critical realism.** *Journal of Management in Medicine* 2000, **14(5/6)**:262-271.
14. Koenig G: **Realistic evaluation and case studies: stretching the potential.** *Evaluation* 2009, **15(1)**:9-30.
15. Chen H-T: *Theory-driven evaluations.* 1 edition. Newbury Park, California: Sage Publications; 1990.
16. Van Der Knaap P: **Theory-based evaluation and learning: possibilities and challenges.** *Evaluation* 2004, **10(1)**:16-34.
17. Donaldson S: **Program theory-driven evaluation science.** Strategies and applications New York: Lawrence Erlbaum Associates; 2007.
18. Stame N: **Theory-based evaluation and types of complexity.** *Evaluation* 2004, **10(1)**:58-76.
19. Dickinson H: **The evaluation of health and social care partnerships: an analysis of approaches and synthesis for the future.** *Health Soc Care Community* 2006, **14(5)**:375-383.
20. Kazi M, Rostila I: **The practice of realist evaluation in two countries.** *European Evaluation Society Conference* Sevilla, Spain; 2002.
21. MacKenzie M, Blamey A: **The practice and theory. Lessons from the application of a Theories of Change approach.** *Evaluation* 2005, **11(2)**:151-168.
22. Leone L: **Realistic evaluation of an illicit drug deterrence programme.** *Evaluation* 2008, **14(1)**:19-28.
23. Ying Ho S: **Evaluating urban regeneration programmes in Britain. Exploring the potential of the realist approach.** *Evaluation* 1999, **5(4)**:422-438.
24. Pommier J, Guével M-R, Jourdan D: **Evaluation of health promotion in schools: a realistic evaluation approach using mixed methods.** *BMC Public Health* 2010, **10(43)**.
25. Guichard A, Ridde V: **Etude exploratoire des mécanismes de l'efficacité des interventions visant à réduire les inégalités sociales de santé.** Etude pilote dans trois régions françaises: Institut National de Prévention et d'Education pour la Santé (INPES), France; 2009.
26. Mays N, Wyke S, D E: **The evaluation of complex health policy.** *Evaluation* 2001, **7(4)**:405-426.
27. Sullivan H, Barnes M, Matka E: **Building collaborative capacity through 'Theories of change'. Early lessons from the evaluation of Health Action Zones in England.** *Evaluation* 2002, **8(2)**:205-226.
28. Byng R: **Using the 'Realistic evaluation' framework to make a retrospective qualitative evaluation of a practice level intervention to improve primary care for patients with long-term mental illness.** *The 2002 EES Conference Three movements in Contemporary Evaluation: Learning, Theory and Evidence October 10-12: 2002* European Evaluation Society; 2002.
29. Wong G, Greenhalgh T, Pawson R: **Internet-based medical education: a realist review of what works, for whom and in what circumstances.** *BMC Med Educ* 2010, **10**:12.
30. Blaise P, Kegels G: **A realistic approach to the evaluation of the quality management movement in health care systems: a comparison between European and African contexts based on Mintzberg's organizational models.** *Int J Health Plann Manage* 2004, **19(4)**:337-364.
31. Marchal B, Dedzo M, Kegels G: **A realist evaluation of the management of a well-performing regional hospital in Ghana.** *BMC Health Services Research* 2010, **10(24)**.
32. Blamey A, Mackenzie M: **Theories of change and realistic evaluation. Peas in a pod or apples and oranges?** *Evaluation* 2007, **13(439-455)**.
33. Pedersen L, Rieper O: **Is realist evaluation a realistic approach for complex reforms?** *Evaluation* 2008, **14(3)**:271-293.
34. Davies R: **Scale, complexity and the representation of theories of change.** *Evaluation* 2004, **10(1)**:101-121.
35. Rogers P: **Using programme theory to evaluate complicated and complex aspects of interventions.** *Evaluation* 2008, **14(1)**:29-48.
36. Lipsey M, Pollard J: **Driving toward theory in program evaluation: more models to choose from.** *Evaluation and Program Planning* 1989, **12**:317-328.
37. Cole G: **Advancing the development and application of theory-based evaluation in the practice of public health.** *American Journal of Evaluation* 1999, **20**:453-470.
38. Yin R: *Case study research. Design and methods.* Third edition. London: Sage Publications; 2003.
39. Weick KE: **Theory construction as disciplined imagination.** *Academy of Management Review* 1989, **14(4)**:516-531.
40. Bourgeois LJ: **Toward a method of middle-range theorizing.** *Academy of Management Review* 1979, **4(3)**:443-447.
41. Campbell C, MacPhail C: **Peer education, gender and the development of critical consciousness: participatory HIV prevention by South African youth.** *Soc Sci Med* 2002, **55(2)**:331-345.
42. Määtä M, Rantala K: **The evaluator as a critical interpreter.** *Evaluation* 2007, **13(4)**:457-476.

Pre-publication history

The pre-publication history for this paper can be accessed here:
http://www.biomedcentral.com/1471-2458/10/741/prepub

doi:10.1186/1471-2458-10-741
Cite this article as: Van Belle *et al.*: How to develop a theory-driven evaluation design? Lessons learned from an adolescent sexual and reproductive health programme in West Africa. *BMC Public Health* 2010 **10**:741.

Northridge and Metcalf *Health Research Policy and Systems* (2016) 14:74
DOI 10.1186/s12961-016-0146-8

Health Research Policy and Systems

REVIEW Open Access

Enhancing implementation science by applying best principles of systems science

Mary E. Northridge[1*] and Sara S. Metcalf[2]

Abstract

Background: Implementation science holds promise for better ensuring that research is translated into evidence-based policy and practice, but interventions often fail or even worsen the problems they are intended to solve due to a lack of understanding of real world structures and dynamic complexity. While systems science alone cannot possibly solve the major challenges in public health, systems-based approaches may contribute to changing the language and methods for conceptualising and acting within complex systems. The overarching goal of this paper is to improve the modelling used in dissemination and implementation research by applying best principles of systems science.

Discussion: Best principles, as distinct from the more customary term 'best practices', are used to underscore the need to extract the core issues from the context in which they are embedded in order to better ensure that they are transferable across settings. Toward meaningfully grappling with the complex and challenging problems faced in adopting and integrating evidence-based health interventions and changing practice patterns within specific settings, we propose and illustrate four best principles derived from our systems science experience: (1) model the problem, not the system; (2) pay attention to what is important, not just what is quantifiable; (3) leverage the utility of models as boundary objects; and (4) adopt a portfolio approach to model building. To improve our mental models of the real world, system scientists have created methodologies such as system dynamics, agent-based modelling, geographic information science and social network simulation. To understand dynamic complexity, we need the ability to simulate. Otherwise, our understanding will be limited. The practice of dynamic systems modelling, as discussed herein, is the art and science of linking system structure to behaviour for the purpose of changing structure to improve behaviour. A useful computer model creates a knowledge repository and a virtual library for internally consistent exploration of alternative assumptions.

Conclusion: Among the benefits of systems modelling are iterative practice, participatory potential and possibility thinking. We trust that the best principles proposed here will resonate with implementation scientists; applying them to the modelling process may abet the translation of research into effective policy and practice.

Keywords: Best principles, Complexity, Context, Implementation science, Modelling, Health equity, Oral health, Primary care, Screening at chairside, Systems science

Background

This review is grounded in the ongoing experiences of the authors to devise and implement interventions to promote health equity, including for older adults. Because the aforementioned interventions are both multilevel and dynamic, the scientific approaches employed evolved from utilising ecological models for thinking through pathways whereby determinants at the societal, community and interpersonal levels affect population and individual health and well-being [1–4], to embracing a portfolio of systems science models that usefully inform related research, practice, policy and education initiatives [5–7].

Forrester, the founder of system dynamics, famously explained that a manager's verbal description of a corporate organisation constitutes a model [8]. Such mental models of corporations are used by managers to deal with problems that arise on a daily basis. They are not, however, the real corporation. Rather, they substitute in our thinking for the real organisation. Sterman, a leading systems scientist modeller and extraordinary communicator, attributes the lack of learning effectively in a world of dynamic complexity to

* Correspondence: men6@nyu.edu
[1]Department of Epidemiology & Health Promotion, New York University College of Dentistry, 433 First Avenue, Room 726, New York, NY 10010, USA
Full list of author information is available at the end of the article

poor inquiry skills. He argues, "*We do not generate alternative explanations or control for confounding variables. Our judgments are strongly affected by the frame in which the information is presented, even when the objective information is unchanged. We suffer from overconfidence in our judgments (underestimating uncertainty), wishful thinking (assessing desired outcomes as more likely than undesired outcomes), and confirmation bias (seeking evidence consistent with our preconceptions)*" ([9], p. 510).

A complex (adaptive) system has been usefully defined as a system comprised of a large number of entities that display a high level of interactivity that is largely nonlinear, containing demonstrable feedback loops [10, 11]. The term systems science is used to refer to the 'big picture' of problem solving, where the problem space is conceptualised as a system of interrelated component parts [12]. Both the coherent whole of the system and the relationships among the component parts are critical to the system, as they give rise to emergence, meaning much coming from little [13]. Note that emergence occurs when even a relatively simple system generates unexpected amounts of complexity, which cannot be understood without the ability to create a model [13]. There are a number of other basic observations that have been made through the examination of complex systems, primarily through the use of computer simulation and the mathematics of nonlinearity, including self-organisation, meaning insensitive to large disturbances [14] and incompressibility, meaning any reduction in complexity will result in the loss of system aspects [15]. The overarching point is that rather than focusing on the parts of a system and how they function, one must focus on the interactions between these parts, and how these relationships determine the identity not only of the parts, but of the whole system [11].

Likewise, dissemination and implementation research places an emphasis on studying issues in context [3, 16, 17]. In his seminal article on diffusion, dissemination and implementation, Lomas explained, "*Implementation … is dependent on a complex framework of sanctions and incentives, reinforced by monitoring and adjustment, and often adapted to fit differing environments at more local levels*" ([18], p. 227). Thus, the congruence of an implementation science approach with a systems science approach is both intuitive and pragmatic. After first-hand engagement in conducting an implementation science pilot study [19, 20], however, the use of systems science modelling to strengthen the dissemination and implementation evidence base became a tangible next step rather than a future direction for the field [21].

Previous researchers have contended systems thinking may usefully advance implementation science. Indeed, Glasgow and Chambers [22] argued that implementation researchers would profit from embracing an interrelated systems perspective rather than a mechanistic, determinism approach to science. Further, Holmes et al. [23] sought to draw attention to certain implications inherent in adopting a systems view for dissemination and implementation research, especially with regard to causation and leverage points for change in a complex system. Recently, Burke et al. [24] presented case examples of three systems science methods, namely system dynamics, agent-based modelling and network analysis, to illustrate how each method may be used to address dissemination and implementation challenges. Finally, Valente conducted a review of network interventions without specifically relating them to implementation science, yet concluded that the choice of intervention depends, in part, on the social context of the program [25], in concert with the systems perspective that context is critical [22].

While complex systems science alone cannot possibly solve the major challenges in public health, it has been argued that systems-based approaches may contribute to changing the language and methods for conceptualising and acting within complex systems [26]. Moreover, it may eventually improve the modelling used in dissemination and implementation research. Toward that end, we thought to share best principles of systems science that we have successfully applied in our own studies toward enhancing implementation science. Best principles, as distinct from the more customary term best practices, are used to underscore the need to extract the core issues from the context in which they are embedded in order to better ensure that they are transferable across settings [27]. For a full treatment of the principles, meaning fundamental truths, of systems science, see the recent text by Mobus and Kalton [28].

The Modelling Process

The problem we were attempting to solve in our pilot study was to improve primary care screening and care coordination at chairside, meaning in a dental setting rather than a medical or other setting [19]. While we had both championed and been involved in previous initiatives that integrated oral health and primary care [29–32], our idea was to support dental hygienists in practicing to the full extent of their training so that they might effectively implement evidence-based guidelines for tobacco use, hypertension and diabetes screening, and nutrition counselling in dental settings [33]. We are principally focused on advancing health equity and ensuring that population groups who lack oral health and primary care are linked to accessible providers and care settings in their own communities, whenever possible [7, 30].

The modelling process is depicted in Fig. 1 as an iterative sequence of steps beginning with problem definition and concluding with policy analysis. Importantly, insights are acquired at all stages of the modelling process.

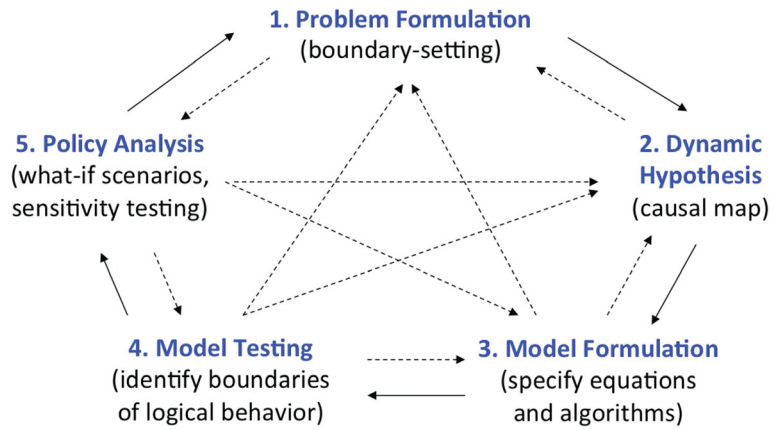


Fig. 1 Stages of the modelling process. The modelling process depicted as an iterative sequence of steps beginning with problem definition and concluding with policy analysis

While Fig. 1 illustrates a return to problem definition upon completion of a modelling project, Sterman [34] emphasises that it may also be appropriate to iterate within the process for the purpose at hand, returning to previous steps or anticipating scenarios ahead of time.

Next, we propose and illustrate four best principles derived from our ongoing systems science research and scholarship that may guide, and perhaps even motivate, implementation scientists in their own studies and thinking. The overarching theme of these best principles involves meaningfully informing the modelling process. It is our belief that this aspect of dissemination and implementation research demands concerted attention in order to meaningfully grapple with the complex and challenging problems faced in adopting and integrating evidence-based health interventions and changing practice patterns within specific settings [31].

Best Principle #1: Model the problem, not the system

Sterman rightly deserves credit for driving home the importance of modelling the problem, not the system [34]. Accordingly, we began our aforementioned pilot study by conducting formative research about the views of dental providers (both dental hygienists and dentists) on primary care coordination at chairside [20]. Findings were that both the dental hygienists and dentists interviewed as part of this research failed to use evidence-based guidelines to screen their patients for primary care-sensitive conditions such as hypertension and diabetes [20]. Nonetheless, all of the participating dental hygienists and dentists reported using electronic devices at chairside to obtain web-based health information in caring for their patients [20]. Hence, we worked collaboratively to develop a clinical decision support system for use by dental hygienists to support them in providing patient care at the level of their full scope of practice [19, 33].

Formerly, we developed a causal map to understand the complex set of causal pathways that are involved and the time delays that accrue over a life course toward developing effective oral health interventions for older adults [5]. A simplified version of this conceptual model is presented below, identifying the key problem variable of our systems science study as "oral health," shown as influencing and influenced by distinct factors at the individual and community scales (Fig. 2). At the individual scale are factors such as nutrition and the presence of chronic illness. Individuals intersect with the community scale in terms of factors such as exposure to oral health promotion interventions and community access to health screening and healthcare.

In subsequent research, we reframed the locus of concern around health equity more broadly, requiring us to reconsider how an individual's health status reflects a broader distribution of social and health disparities that vary by population subgroups. An orientation toward health equity warrants a broader model conceptualisation than health per se [35].

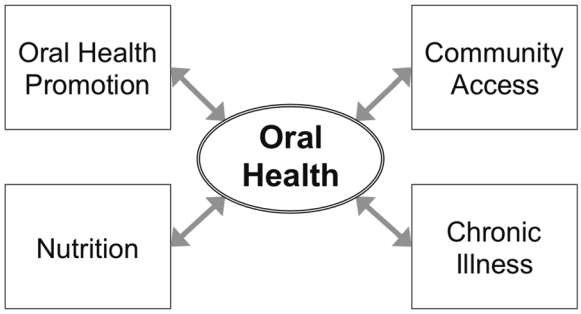


Fig. 2 Conceptual model focused on the problem of oral health. A conceptual model that identifies the key problem variable of a systems science study as oral health, shown as influencing and influenced by distinct factors at the individual and community scales

Attempts to model the system rather than the problem are bound to lead to confusion and futility [34]. Our training and experience in systems science directed us away from trying to design an integrated system of oral and primary care and focused our attention instead on supporting dental hygienists to adhere to evidence-based tobacco use, hypertension and diabetes screening, and nutrition counselling referral guidelines [19, 20, 33]. Formative research and interdisciplinary collaboration are invaluable in steering implementation scientists toward modelling the problem, not the system.

Best Principle #2: Pay attention to what is important, not just what is quantifiable

Meadows was a rigorous systems scientist who inspired her colleagues and students to pay attention to what is important – be it justice, democracy, security, freedom, truth, or love – even if it cannot be precisely defined or measured [36]. Unfortunately, despite the critical importance of qualitative information, certain researchers restrict the constructs and variables in their models to those for which numerical data are available, and include only those parameters that can be estimated statistically [37]. Yet, in a comprehensive article on collecting and analysing qualitative data for system dynamics [38], Luna-Reyes and Andersen argue convincingly that qualitative data and their analysis also have a central role to play at all stages of the modelling process. Using strategies such as theirs, qualitative statements can be used to derive causal relationships.

As an example, in a Spanish-language focus group about dental care conducted with men aged 50 years and older who reside in northern Manhattan, New York, and had immigrated from the Dominican Republic, one participant explained: "*Sometimes you* [go to the dentist] *because you get a referral from a friend: 'Oh, so and so. Now that's a good dentist.' So you go, more or less, because of that reference. It's not like you go* [because of] *where it is, but because you had a referral, and that information circulates.*" This explanation summarises the importance of the peer network in recommending healthcare providers. A reinforcing loop reflecting the essence of this comment is depicted in Fig. 3. The notion that information circulates points to the mechanism by which an individual's experience with a provider translates into referrals or recommendations for the provider, inducing her or his social ties to then pursue care with the recommended provider. An intermediate construct of trust in healthcare provider extends beyond the direct comment but helps to articulate the basis of the recommendation.

Because dissemination and implementation studies are based on the mechanisms through which health information, interventions and evidence-based clinical practices are adopted in public health, community and healthcare service use in a variety of settings, a broad range of methodological approaches are employed [39]. These include both traditional designs, such as randomised controlled trials, and newer approaches such as hybrid effectiveness-implementation designs [40, 41]. While mixed methods approaches are endorsed in implementation science, there is a need for greater attention to connectedness across program levels and components [40].

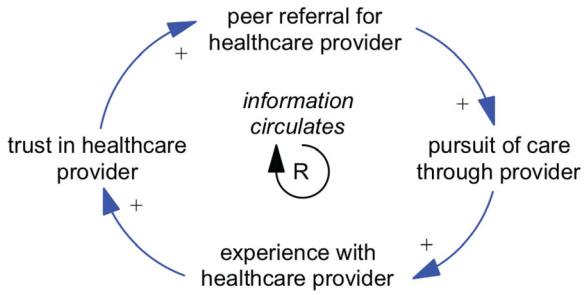


Fig. 3 Example of causal mapping from qualitative data. A reinforcing loop reflecting the essence of a comment (qualitative data) from a focus group participant regarding the importance of a peer network in recommending healthcare providers

We are at the point in our implementation science study of primary care coordination by dental hygienists at chairside where we need to create a causal map (also known as causal loop diagram) to provide a systematic way to develop dynamic hypotheses and identify important feedback loops [42]. In a causal map, it is possible to ascribe certain variables to specific scales, e.g. community, interpersonal and individual. Because systems science models are not limited to constructs that are precisely defined or measured, deep thinking and multiple perspectives may help guide implementation scientists to pay attention to what is important, not just what is quantifiable.

Best Principle #3: Leverage the utility of models as boundary objects

According to Black, a boundary object is "*a representation—perhaps a diagram, sketch, sparse text, or prototype—that helps individuals collaborate effectively across some boundary, often a difference in knowledge, training, or objective*" ([43], p. 76). For research teams such as ours, whose members possess expertise in diverse domains, boundary objects are useful for coordinating knowledge and objectives and for developing a shared vocabulary about the problem to be solved collaboratively [44].

The conceptual framework that informs our interventions is the Consolidated Framework for Implementation Research (CFIR) [45]. While this proved to be incredibly helpful to us in designing and evaluating our implementation science pilot study, we found the accompanying graphic to be difficult to understand.

Hence, we developed a simplified model that was derived from previous examples used in our systems science research. As shown in Fig. 4, the five major domains of the CFIR (the intervention, the inner setting, the outer setting, the individuals involved and the process by which implementation is accomplished) are represented in the simplified model, along with the process of adaptation [20].

This graphic proved to be both intuitive and accessible to our interdisciplinary team members, so much so that we have created project-specific models for a series of papers [19, 20, 33]. We now consider our CFIR model to be a boundary object that facilitates team collaboration.

Note that, from a modelling perspective, a boundary object is "*a socially constructed artefact for building trust and agreement*" ([46], p. 4, citing [47]). For boundary objects to be useful, they must be modifiable and readily perceptible representations that embody the dependencies among resources and goals of team members [48]. While boundary objects represent local knowledge, they may be shared across networks and thus play a significant role in creating synergies which in turn sustain local initiatives [49]. Developed models used as boundary objects may benefit implementation scientists through building trust and agreement that represent local knowledge.

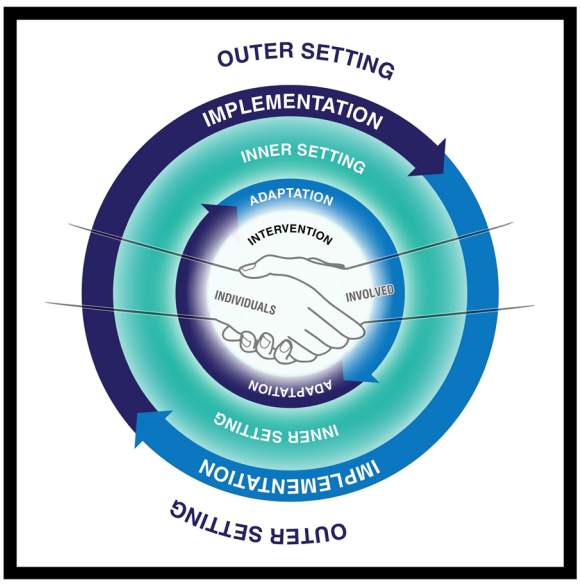


Fig. 4 Simplified model of the Consolidated Framework for Implementation Research. The five major domains of the Consolidated Framework for Implementation Research (the intervention, the inner and outer settings, the individuals involved, and the process by which implementation is accomplished) are represented in this simplified model, along with the process of adaptation

Best Principle #4: Adopt a portfolio approach to model building

As we alluded to at the outset of this paper, our research team led by the authors – an implementation scientist (MEN) and a systems scientist (SSM) – has developed a portfolio of conceptual, statistical, spatial and simulation models that utilise the multiple information streams associated with our research projects [44]. A chief advantage of the portfolio approach in a collaborative research context is that it provides multiple entry points and checkpoints to the modelling process, facilitating input from different team members [6]. A further benefit is that team members often work in parallel to develop separate but related models in diverse ways, exploring the simulated consequences of alternative assumptions [6].

For instance, in our ongoing project, Integrating Social and Systems Science Approaches to Promote Oral Health Equity, our modelling team has gained important insights by adopting a portfolio approach that incorporates different methods of systems science, including system dynamics, agent-based modelling, geographic information science and social network simulation, in models that help to explore challenges to realising oral health equity for older adults [6, 35]. This portfolio approach to systems science modelling enables our research team to interpret and triangulate between different scenarios at distinct geographic and temporal scales. An inventory of the simulation models in our portfolio that highlights their links to other models in the portfolio is provided in Additional file 1.

In essence, then, the construction of a portfolio of models confers flexibility to the modelling process and is especially conducive to collaboration, allowing for multiple opportunities for input and adjustment of models by different members of the research team. Further, the portfolio approach leverages the iterative nature of the modelling process and encourages exploration with 'flawed' models rather than

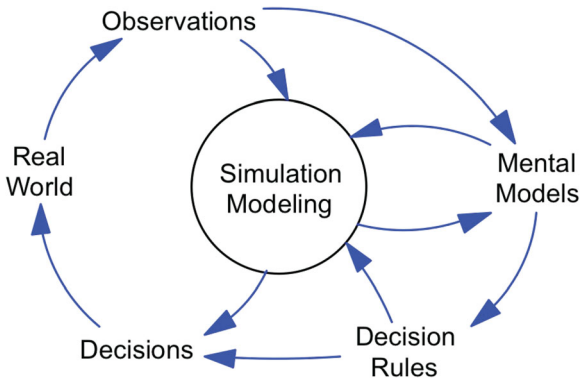


Fig. 5 Simulation modelling in context. The practice of simulation modelling is situated amidst an ongoing process of observing the real world, formulating mental models of how it works, setting decision rules to guide behaviour, and from these heuristics, making decisions that in turn affect the state of the real world

Table 1 Summary of best principles from systems science for informing the modelling process, recommendations for action by implementation scientists and contributing thought leaders and key references

Best principle	Recommendations	Thought leader [Reference]
1. Model the problem, not the system	Conduct formative research; construct models collaboratively in interdisciplinary teams	Sterman [34]
2. Pay attention to what is important, not just what is quantifiable	Use qualitative data to derive causal relationships; be guided by deep thinking and multiple perspectives	Meadows [36]
3. Leverage the utility of models as boundary objects	Create modifiable and readily perceptible representations of models; build trust and agreement by representing local knowledge	Black [43]
4. Adopt a portfolio approach to model building	Work in parallel to develop separate but related models in diverse ways; encourage exploration with 'flawed' models rather than aiming for perfection	Metcalf [6]

aiming for perfection with 'kitchen sink' models. Implementation scientists may profit from adopting a portfolio approach to model building that confers flexibility and is conducive to collaboration.

Conclusions

In order to improve our mental models of the real world, system scientists have developed and leveraged methods such as system dynamics, agent-based modelling, geographic information science and social network simulation. As articulated by Sterman [34] (Fig. 5), the practice of simulation modelling is situated amidst an ongoing process of observing the real world, formulating mental models of how it works, setting decision rules to guide behaviour, and from these heuristics, making decisions that in turn affect the state of the real world. Simulation modelling offers a mechanism for what Sterman calls 'double-loop learning' [34], arriving at insight from the process of virtual experimentation afforded by simulation modelling, in addition to learning from experiences in the real world. The two-way relationship between mental models and simulation modelling underscores the essential nature of learning through the modelling process.

Because as humans we can only process a limited amount of information in our heads as 'thought experiments,' we need to simulate computer models to transcend our mental models. In short, to understand dynamic complexity, we need the ability to simulate. Otherwise, our understanding will be limited.

Modelling, then, is the art and science of linking system structure to behaviour for the purpose of changing structure to improve behaviour. A useful computer model creates a knowledge repository and a virtual library for internally consistent exploration of alternative assumptions. Among the benefits of systems modelling are iterative practice, participatory potential and possibility thinking.

We trust that the best principles proposed here will resonate with our fellow implementation scientists and that applying them to the modelling process will abet the translation of research into effective policy and practice. Table 1 provides a summary of the four best principles discussed herein for informing the modelling process, along with recommendations for action by implementation scientists and the contributing thought leaders whose references we cited.

As Sterman cautions us, "*What prevents us from overcoming policy resistance is not a lack of resources, technical knowledge, or a genuine commitment to change. What thwarts us is our lack of a meaningful systems thinking capability*" ([9], p. 513).

Additional file

Additional file 1: Summary of simulation models in systems science portfolio. (DOCX 22 kb)

Funding

The authors were supported in the research, analysis and writing of this paper by the National Center for Advancing Translational Sciences of the US National Institutes of Health for the project entitled, *Primary Care Screening by Dental Hygienists at Chairside: Developing and Evaluating an Electronic Tool* (grant UL1TR000038) and by the National Institute for Dental and Craniofacial Research and the Office of Behavioral and Social Sciences Research of the US National Institutes of Health for the project entitled, *Integrating Social and Systems Science Approaches to Promote Oral Health Equity* (grant R01-DE023072).

Authors' contributions

MEN conceived of the study, participated in its design, wrote the first draft, and contributed to the conceptualisation of the figures. SSM participated in the design of the study, provided substantive edits to the draft and created the figures. Both authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

All Columbia University, New York University, and University at Buffalo institutional review board and Health Insurance Portability and Accountability Act safeguards were followed.

Author details
[1]Department of Epidemiology & Health Promotion, New York University College of Dentistry, 433 First Avenue, Room 726, New York, NY 10010, USA. [2]Department of Geography, The State University of New York at Buffalo, 115 Wilkeson Quad, Ellicott Complex, North Campus, Buffalo, NY 14261, USA.

Received: 4 July 2016 Accepted: 20 September 2016
Published online: 04 October 2016

References

1. Northridge ME, Sclar ED, Biswas P. Sorting out the connections between the built environment and health: a conceptual framework for navigating pathways and planning healthy cities. J Urban Health. 2003; 80(4):556–68.
2. Schulz A, Northridge ME. Social determinants of health: implications for environmental health promotion. Health Educ Behav. 2004;31(4):455–71.
3. Glass TA, McAtee MJ. Behavioral science at the crossroads in public health: extending horizons, envisioning the future. Soc Sci Med. 2006;62(7):1650–71.
4. Northridge ME, Ue F, Borrell LN, Bodnar S, De La Cruz L, Marshall S, Lamster IB. Tooth loss and dental caries in community-dwelling older adults in northern Manhattan. Gerodontology. 2012;29(2):e464–73.
5. Metcalf SS, Northridge ME, Lamster IB. A systems perspective for dental health in older adults. Am J Public Health. 2011;101(10):1820–3.
6. Metcalf SS, Northridge ME, Widener MJ, Chakraborty B, Marshall SE, Lamster IB. Modeling social dimensions of oral health among older adults in urban environments. Health Educ Behav. 2013;40(1S):63S–73.
7. Metcalf SS, Birenz SS, Kunzel C, Wang H, Schrimshaw EW, Marshall SE, Northridge ME. The impact of Medicaid expansion on oral health equity for older adults: a systems perspective. J Calif Dent Assoc. 2015;43(7):379–87.
8. Forrester JW. Industrial Dynamics. Waltham: Pegasus Communications; 1961.
9. Sterman JD. Learning from evidence in a complex world. Am J Public Health. 2006;96(3):505–14.
10. Richardson K, Cilliers P. What is complexity science? A view from different directions. Emergence. 2001;3(1):5–23.
11. Richardson KA, Cilliers P, Lissack M. Complexity science: a "gray" science for the "stuff in between". Emergence. 2001;3(2):6–18.
12. Mabry PL, Olster DH, Morgan GD, Abrams DB. Interdisciplinarity and systems science to improve population health: a view from the NOH Office of Behavioral and Social Sciences Research. Am J Prev Med. 2008;35(2S):S211–24.
13. Holland JH. Emergence: From Chaos to Order. New York: Basic Books; 1998.
14. Auyang SY. Foundations of Complex-System Theories in Economics, Evolutionary Biology, and Statistical Physics. Cambridge: Cambridge University Press; 1999.
15. Cilliers P. Complexity and Postmodernism: Understanding Complex Systems. London: Routledge; 1998.
16. Biglan A. Changing Cultural Practices: A Contextualist Framework for Intervention Research. Reno: Context Press; 1995.
17. Stokols D, Misra S, Moser RP, Hall KL, Taylor BK. The ecology of team science: understanding contextual influences on transdisciplinary collaboration. Am J Prev Med. 2008;35(2 Suppl):S96–115.
18. Lomas J. Diffusion, dissemination, and implementation: who should do what? Ann N Y Acad Sci. 1993;703:226–35. Discussion 235–7.
19. Russell SL, Greenblatt AP, Gomes D, Birenz S, Golembeski CA, Shelley D, McGuirk M, Eisenberg E, Northridge ME. Toward implementing primary care at chairside: developing a clinical decision support system for dental hygienists. J Evid Based Dent Pract. 2015;15(4):145–51.
20. Northridge ME, Birenz S, Gomes G, Golembeski CA, Greenblatt AP, Shelley D, Russell SL. Views of dental providers on primary care coordination at chairside. J Dent Hyg. 2016;90(3):195–205.
21. Glasgow RE, Vinson C, Chambers D, Khoury MJ, Kaplan RM, Hunter C. National Institutes of Health approaches to dissemination and implementation science: current and future directions. Am J Public Health. 2012;102(7):1274–81.
22. Glasgow RE, Chambers D. Developing robust, sustainable, implementation systems using rigorous, rapid and relevant science. Clin Transl Sci. 2012;5(1):48–55.
23. Holmes BH, Finegood DT, Riley BL, Best A. Systems thinking in dissemination and implementation research. In: Brownson RC, Colditz GA, Proctor EK, editors. Dissemination and Implementation Research in Health: Translating Science to Practice. New York: Oxford University Press; 2012. p. 175–91.
24. Burke JG, Lich KH, Neal JW, Meissner HI, Yonas M, Mabry PL. Enhancing dissemination and implementation research using systems science methods. Int J Behav Med. 2015;22(3):283–91.
25. Valente TW. Network interventions. Science. 2012;337(6090):49–53.
26. Carey G, Malbon E, Carey N, Joyce A, Crammond B, Carey A. Systems science and systems thinking for public health: a systematic review of the field. BMJ Open. 2015;5(12):e009002.
27. Sclar ED, Northridge ME, Karpel EM. Promoting interdisciplinary curricula and training in transportation, land use, physical activity, and health. In: Does the Built Environment Influence Physical Activity? Examining the Evidence. Transportation Research Board Special Report 282. Washington, DC: Transportation Research Board; 2005.
28. Mobus GE, Kalton MC. Principles of Systems Science. New York: Springer; 2015.
29. Northridge ME, Glick M, Metcalf SS, Shelley D. Public health support for the health home model. Am J Public Health. 2011;101(10):1818–20.
30. Northridge ME, Yu C, Chakraborty B, Port A, Mark J, Golembeski C, Cheng B, Kunzel C, Metcalf SS, Marshall SE, Lamster IB. A community-based oral public health approach to promote health equity. Am J Public Health. 2015; 105 Suppl 3:S459–65.
31. Marshall SE, Cheng B, Northridge ME, Kunzel C, Huang C, Lamster IB. Integrating oral and general health screening at senior centers for minority elders. Am J Public Health. 2013;103(6):1022–5.
32. Marshall SE, Schrimshaw EW, Kunzel C, Metcalf SS, Greenblatt AP, De La Cruz LD, Northridge ME. Evidence from ElderSmile for diabetes and hypertension screening in oral health programs. J Calif Dent Assoc. 2015; 43(7):379–87.
33. Westphal Theile C, Strauss SM, Northridge ME, Birenz S. The oral health care manager in a patient-centered health facility. J Evid Based Dent Pract. 2016; 16(Suppl):34–42.
34. Sterman JD. Business Dynamics: Systems Thinking and Modeling for a Complex World. New York: The McGraw-Hill Companies, Inc.; 2000.
35. Metcalf SS, Northridge ME. Engaging in systems science to promote health equity. SAGE Research Methods Case. (in press)
36. Meadows DH. Thinking in Systems: A Primer. White River Junction: Chelsea Green Publishing Co.; 2008.
37. Sterman JD. All models are wrong: reflections on becoming a systems scientist. Syst Dyn Rev. 2002;18:501–31.
38. Luna-Reyes LF, Andersen DL. Collecting and analyzing qualitative data for system dynamics: methods and models. Syst Dyn Rev. 2003;19:271–96.
39. Brownson RC, Colditz GA, Proctor EK, editors. Dissemination and Implementation Research in Health: Translating Science to Practice. New York: Oxford University Press; 2012.
40. Glasgow RE, Emmons KM. How can we increase translation of research into practice? Types of evidence needed. Ann Rev Public Health. 2007; 28:413–33.
41. Curran GM, Bauer M, Mittman B, Pyne JM, Stetler C. Effectiveness-implementation hybrid designs: combining elements of clinical effectiveness and implementation research to enhance public health impact. Med Care. 2012;50(3):217–26.
42. Metcalf SS, Kum SS. System dynamics. In: Richardson D, Castree N, Goodchild MF, Kobayashi A, Liu W, Marston RA, editors. International Encyclopedia of Geography: People, the Earth, Environment, and Technology. Hoboken: Wiley-Blackwell and the Association of American Geographers; 2016.
43. Black LJ. When visuals are boundary objects in system dynamics work. Syst Dyn Rev. 2013;29(2):70–86.
44. Kum SS, Wang H, Jin Z, Xu W, Mark J, Northridge ME, Kunzel C, Marshall SE, Metcalf SS. Boundary objects for group model building to explore oral health equity. Cambridge: 33rd International Conference of the System Dynamics Society; 2015. http://www.systemdynamics.org/conferences/2015/papers/P1302.pdf. Accessed 3 Mar 2016.
45. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implement Sci. 2009;4:50.
46. Scott RJ, Cavana RY, Cameron D. Mechanisms for understanding mental model change in group model building. Syst Res Behav Sci. 2016;33(1):100–18.
47. Zagonel AA. Model conceptualization in group model building: a review of the literature exploring the tension between representing reality and negotiating a social order. Proceedings of the 20th International System Dynamics Conference. Palermo: System Dynamics

Society; 2002. http://www.systemdynamics.org/conferences/2002/proceed/papers/Zagonel1.pdf. Accessed 3 Mar 2016.
48. Black LJ, Andersen DF. Using visual representations as boundary objects to resolve conflict in collaborative model-building approaches. Syst Res Behav Sci. 2012;29(2):194–208.
49. Nyella E, Nguyen T, Braa J. Collaborative knowledge making and sharing across sites: the role of boundary objects. Mediterranean Conference on Information Systems (MCIS) 2010 Proceedings. Paper 64. http://aisel.aisnet.org/mcis2010/64. Accessed 3 Mar 2016.

Liu et al. Systematic Reviews (2016) 5:138
DOI 10.1186/s13643-016-0314-5

Systematic Reviews

PROTOCOL Open Access

Examining the use of process evaluations of randomised controlled trials of complex interventions addressing chronic disease in primary health care—a systematic review protocol

Hueiming Liu*, Janini Muhunthan, Adina Hayek, Maree Hackett, Tracey-Lea Laba, David Peiris and Stephen Jan

Abstract

Background: Randomised controlled trials (RCTs) of complex interventions in primary health care (PHC) are needed to provide evidence-based programmes to achieve the Declaration of Alma Ata goal of making PHC equitable, accessible and universal and to effectively address the rising burden from chronic disease. Process evaluations of these RCTs can provide insight into the causal mechanisms of complex interventions, the contextual factors, and inform as to whether an intervention is ineffective due to implementation failure or failure of the intervention itself. To build on this emerging body of work, we aim to consolidate the methodology and methods from process evaluations of complex interventions in PHC and their findings of facilitators and barriers to intervention implementation in this important area of health service delivery.

Methods: Systematic review of process evaluations of randomised controlled trials of complex interventions which address prevalent major chronic diseases in PHC settings. Published process evaluations of RCTs will be identified through database and clinical trial registry searches and contact with authors. Data from each study will be extracted by two reviewers using standardised forms. Data extracted include descriptive items about (1) the RCT, (2) about the process evaluations (such as methods, theories, risk of bias, analysis of process and outcome data, strengths and limitations) and (3) any stated barriers and facilitators to conducting complex interventions. A narrative synthesis of the findings will be presented.

Discussion: Process evaluation findings are valuable in determining whether a complex intervention should be scaled up or modified for other contexts. Publishing this protocol serves to encourage transparency in the reporting of our synthesis of current literature on how process evaluations have been conducted thus far and a deeper understanding of potential challenges and solutions to aid in the implementation of effective interventions in PHC beyond the research setting.

Systematic review registration: PROSPERO CRD42016035572

Keywords: Process evaluations, Primary health care, Complex interventions, Systematic review, Chronic disease, Qualitative

* Correspondence: hliu@georgeinstitute.org.au
The George Institute for Global Health, University of Sydney, Level 10, King George V Building, 83-117 Missenden Rd, PO Box M201, Camperdown, NSW 2050, Australia

Background

Why is this field of research important?

With a rapidly rising global burden of disease attributed to non-communicable diseases, access to high quality primary health care (PHC) is essential. Complex interventions, defined as '*interventions that comprise multiple interacting components, although additional dimensions of complexity include the difficulty of their implementation and the number of organisational levels they target*', are frequently deployed in an attempt to address health system deficiencies experienced by patients and providers [1]. Choosing a study design to assess effectiveness of complex interventions is not straightforward, and it is recommended to consider randomisation to prevent selection bias and provide robust evidence [2, 3]. Process evaluations, which are typically carried out in conjunction with randomised controlled trials of such interventions, can help explain for whom, how and why a complex intervention had a particular impact [4].

Such evaluations address the question 'Is this intervention acceptable, effective, affordable and feasible (for me or) for this population?' [5]. Process evaluations can enable patient-centred care by providing the opportunity for often over-looked patients' perspectives to be considered. As an example, while a pragmatic trial of a cardiovascular polypill in Australian PHC indicated the polypill was an effective, cost-effective strategy for improving patient adherence and the prescribing of indicated medications, our process evaluation interviews found that clinicians need to consider the polypill strategy alongside other evidence-based strategies. These strategies should cater to specific patient factors such as health literacy, sense of well-being, financial considerations, establishing ongoing respectful clinician and patient relationships and improving accessibility to health care [6].

Despite the generation of good quality evidence, this often does not translate into improved health outcomes [7]. A key barrier in the literature to research translation is cost at different levels, e.g. high outpatient costs for screening to the patient or cost of medications for the programme [8–10]. While health economic evaluations are increasingly being conducted as separate studies to provide evidence of cost-effectiveness to decision makers, there may be cost information that is relevant to the objectives of a process which needs to be investigated. For example, minimising indirect costs to patients is something that is important in understanding why an intervention may be more acceptable to patients compared to standard care. Conversely, for some, indirect costs associated with the intervention may discourage patients from seeking care. These are economic issues which we propose would be important to capture as part of process evaluations but are not strictly captured in health economic evaluations assessing the incremental cost-effectiveness of interventions. It would be pertinent as part of a process evaluation to incorporate relevant cost data from the onset, especially within PHC trials in low- and middle-income countries (LMIC) and in populations which have complex needs and limited funding to be allocated [9]. This would be important as part of a process evaluation, to unpack whether for whom and how an intervention can be implemented into routine practice after the trial is completed. These findings from process evaluations can then inform adoption of interventions into practice and thus the scalability and sustainability of interventions [11].

What is known about this field currently?

Process evaluation methodology is evolving [4]. Process evaluations were previously synonymous with qualitative research alongside trials and were conducted to provide a deeper understanding of the disease condition, implementation issues and mechanisms of the intervention [12]. However, there is a growing recognition that using qualitative and quantitative data (mixed methods) can help facilitate trial implementation and research translation [13–15]. For instance, stratifying quantitative outcome data by socio-economic status and triangulating it with qualitative interviews, multi-level modelling and embedded cost-analysis in a process evaluation may be useful in determining the relevance and feasibility of a proven effective complex intervention. Using mixed methods, a clearer picture of the intervention may emerge that could aid various stakeholders in their decision-making.

Although 'one size fits all' methods or methodologies are not available, various theories or frameworks to enhance implementation research have been used by researchers to assist in their process evaluations. In early 2015, guidance was published by the Medical Research Council (MRC) UK about the planning, conduct and reporting of process evaluations to aid researchers, policy makers and funders [11]. The article described the proposed functions of process evaluations of looking at feasibility and piloting, evaluation of effectiveness and implementation post-evaluation during the different stages of the development, evaluation and implementation of a complex intervention. These functions expanded upon the conventional definition of process evaluations being limited to during trial implementation and defined 'implementation' as '*the process through which interventions are delivered, and what is delivered in practice*'. For example, during the post-evaluation implementation stage, the authors recommend that the process evaluation serves to explore how there is '*routinisation of the intervention into new contexts, and long term implementation/maintenance*'. The authors suggest that this function of the process evaluation is needed as reviews have showed that post-trial, complex interventions are only partially maintained. Key recommendations regarding the planning,

design and conduct, analysis and reporting of process evaluations were also discussed in the MRC recommendations [4, 11]. For example, arguments for whether there should be a separation or integration of the process evaluation and outcome evaluation teams were presented. The need to integrate process and outcome data in the analysis and the timing of when process data should be analysed in relation to outcome data were discussed.

The appraisal of the quality of process evaluations has not been straightforward partly because of the variability in methods [11, 16, 17]. Grant et al. in a literature review found that the process evaluations were of poor and inconsistent quality and proposed seven criteria for the reporting of process evaluations including clearly labelling that it is a process evaluation [17]. Other suggestions include appraising the quality of the process evaluation based on the methods used. Given that most process evaluations will have a qualitative component, a set of criteria to examine the quality in the reporting of qualitative research will be relevant to most process evaluations [18, 19].

Dissemination and reporting of findings from process evaluations especially in academic publications can also be difficult due to a variety of reasons including feasibility due to limited resources for research projects, lag time till dissemination of result or publication bias as usually positive outcome trials will be reported but not necessary negative trials [11, 20]. This in turn could limit the likelihood of such relevant findings affecting policy and practice.

Why do this review?

The George Institute for Global Health has a current programme of research which focuses on addressing NCDs through cost-effective and equitable strategies in primary health care settings including LMIC, and with indigenous populations [21]. Our studies trial complex interventions such as capacity-building initiatives with local providers [22], use of innovative mobile technology [23], and cost-effective generic medications (e.g. polypill) within primary health care settings [24]. We have found that at times, despite acceptability and effectiveness of these strategies, there are significant challenges that impact upon their scale up. These barriers could be cultural, political or institutional factors [25], but an important reason for limited translation seems to lie in the lack of understanding of implementation issues within contextual factors for the different stakeholders (e.g. patient, provider, policy makers). For example, while a trial in India of a clinical decision support system on a mobile tablet improved initial diagnosis and antihypertensive management of trial patients, and was deemed acceptable by end-users, only 35% of patients attended the scheduled 1-month follow-up [23]. Interviews with stakeholders found that limited patient accessibility to medicines and doctors (for a variety of reasons including inadequate staffing, limited primary health care infrastructure) as the key barrier which needs to be overcome. This contrasts with other trials of electronic health tools (e.g. decision support, text messages) in Australia which tend towards generally more positive and sustained results as such presumably because system issues were less of a significant barrier given the universal and subsidised health care available [26–28]. Given the greater burden of early mortality from NCD in LMIC and disadvantaged populations [29], consolidating our findings in this proposed systematic review with an equity-focused lens to better understand how to strengthen PHC within relevant contextual policy and system issues would be useful. Indeed, systematic reviews of interventions in primary health care have concluded that in addition to clinical outcomes, rigorous evaluations of implementation outcomes (e.g. through process evaluations) are needed to ensure changes in practice [30, 31]. We hope that this systematic review will add to the process evaluation methodology and understanding of effective implementation strategies in different PHC settings [32, 33].

Objectives and key questions

To our knowledge, this is the first systematic review of process evaluations of randomised controlled trial (RCTs) in complex interventions in PHC. For complex interventions, the pre-specification of a theory for how an intervention is expected to work can be highly informative in identifying the mechanisms by which an intervention was hypothesised to have an impact and why it was found to be successful (or not). It provides a framework for assessing the behaviour of individual actors in the implementation of an intervention, potential breakdowns in the interactions between parties and puts into context these actions. Thus, findings from process evaluations from both positive and negative trials can shed light upon implementation facilitators and barriers, which would add to the collective lessons for researchers. Moreover, given that there are numerous theories and frameworks in this area, we thought it would be informative to describe the breadth of methods used and to make some recommendations on evaluation methods that should be incorporated into PEs of complex interventions. Thus, we aim to consolidate the methodology and methods from process evaluations of complex interventions in PHC and their findings of facilitators and barriers to intervention implementation in this important area of health service delivery.

These objectives will be achieved through addressing these questions: (a) Is there and what is the explicit theory behind the conducted process evaluations? (e.g. normalisation process theory, realist framework); (b) What are

the methods used in these process evaluations? (e.g. qualitative research through semi-structured interviews, surveys); (c) At what stage is the process evaluation done? (i.e. feasibility and piloting, evaluation of effectiveness, or post-evaluation implementation.); (d) If an aim is stated (i.e. in the evaluation of effectiveness stage), how are the results of the RCT integrated with the findings from the process evaluations?; (e) What are the strengths, limitations and potential solutions identified by the authors in conducting the process evaluations?; and (f) What are the barriers and facilitators to the implementation of complex interventions identified by the authors?

Methods/design

This systematic review will focus on process evaluations of RCTs of complex interventions addressing chronic disease in PHC. We have described our methods as per Preferred Reporting Items for Systematic Review and Meta-Analysis for protocol (PRISMA-P) recommendations, and this checklist is included as an Additional file 1 [34].

Eligibility criteria

Definitions as per PICO-D have been adapted for the purpose of this review [20, 35]:

Participants—participants include patients and health providers in the PHC setting addressing the prevalent chronic diseases as defined by the World Health Organisation—cardiovascular disease, chronic kidney disease, chronic respiratory disease, type 2 diabetes mellitus and depression. PHC as defined by the Alma-Ata declaration [36] as health services provided within the community setting by doctors, nurses and allied health with the goal to achieve better health for all through reforms in universal coverage, public policy, service delivery and leadership [37].

'Intervention'—complex interventions defined as those '*interventions that comprise multiple interacting components, although additional dimensions of complexity include the difficulty of their implementation and the number of organisational levels they target*' within PHC [4]. This includes a single-faceted intervention that requires multiple actors or pathways and thus makes the implementation complex. It is envisaged that the complex interventions for chronic diseases (if not explicitly defined as a complex intervention) will have elements of the Wagner chronic care model such as community support, case management, self-management, facilitated family support, organisational change, delivery system design, decision support for health care providers and clinical information systems [38].

Comparator—not applicable

'Outcomes'—(1) findings from the process evaluations of stated implementation barriers and facilitators to the complex intervention. (2) The stated strengths and limitations of the process evaluation methodology from the perspectives of the authors. Both findings will be useful for future conduct of complex interventions in PHC in the planning, conduct of process evaluations and in the consideration of intervention implementation and what barriers need to be overcome in different PHC settings.

Timing—years of search from 1998. This was chosen because a systematic review by Davies et al. shows that there was poor use of theory in implementation research until at least 1998 [39].

Design—process evaluations of randomised controlled trials of complex interventions in PHC. Process evaluation as defined by 'a study which aims to understand the functioning of an intervention, by examining implementation, mechanisms of impact, and contextual factors' [11]. As discussed by Grant et al., because process evaluations are not clearly labelled as such, qualitative research conducted alongside RCTs with similar aims will be included [17, 40].

Exclusion criteria—articles were excluded if they were not a journal article, not a report based on empirical research (e.g. protocol, editorial), not reported in English and reviews and not human research.

Search strategy

Information sources

Databases reporting academic publications (MEDLINE, SCOPUS, PsychInfo, CINAHL, EMBASE, Global Health.) In order to locate any process evaluations whose findings were not published or missed in the database searches, we will search major clinical trial registries for completed process evaluations (e.g. Cochrane Central Registry of Controlled Trials, EU registry, ANZTRN and clinical trial registry (USA)). Authors will be contacted in regard to the outcomes of the RCT and findings of their completed process evaluations.

A search strategy was developed and adapted for each database with the initial support of a medical research librarian. Search terms were based on the review objectives and early scoping searches (see Additional file 2: search strategy), key words: process evaluations (including programme evaluation, qualitative research), complex intervention (including chronic care model and its components of community support, case management, self-management, facilitated family support, organisational change, delivery system design, decision support for health care providers and clinical information systems), randomised controlled trials, PHC (including family practice, general practitioners) and chronic disease (including cardiovascular disease, chronic kidney disease, chronic respiratory disease, type 2 diabetes mellitus and depression).

Study records

Data management

After the searches, the shortlisted articles will be exported to Endnote. Data will be stored in a common file that is

password protected on the Institute's server that is accessible by the two reviewers. At each stage of the data selection process during the review (e.g. after consolidation of all articles prior to assessing eligibility based on title and abstract), back up files of the endnote database will be made in order to retrace any steps as needed in the review process, and for any third party adjudication.

Selection process

Two reviewers will screen all titles and abstracts identifying potential eligible studies based on inclusion and exclusion criteria, and duplicates are to be removed. This will be done independently to reduce the risk of bias. All eligible studies will be retrieved in full text and reviewed by the two reviewers using predesigned eligibility forms (see Additional file 3). Disagreements will be resolved by consensus of a third party in the review team.

Data collection process

Data from all included studies will be extracted by two reviewers using the eligibility and data extraction forms. The data extraction forms (see Additional file 4) were partly guided by the MRC recommendation for process evaluations and Grant et al.'s suggested minimal factors for reporting on process evaluations [4, 17]. The forms will be pilot tested by the two reviewers on the same three articles, iterative changes will be made when appropriate and the two reviewers will independently extract data from the rest of the included list of articles.

Data items

Variables to be extracted include data on the RCT and its process evaluation: (1) RCT—study design, setting (rural, urban, country), results (positive, negative or equivalent); (2) process evaluation—any published process evaluation protocol or evidence of pre-specified process evaluation in the main trial protocol, or stated aims of the process evaluation (e.g. examining recruitment, or explaining results), the process evaluation theory, justified methods of integrating trial and process outcomes, stage during which the process evaluation is done (feasibility and piloting, evaluation of effectiveness and post-evaluation implementation), methods of analysis and inclusion of costs incurred.

Outcomes and prioritisation

The outcomes of interest for our aims are (1) the stated strengths and limitations of the process evaluation methodology from the perspectives of the authors and (2) findings from the process evaluation of stated implementation barriers and facilitators of the complex intervention. Both findings will be useful for future conduct of complex interventions in PHC—in the planning, conduct of process evaluations and when considering the scaling up of an interventions and what barriers need to be overcome in a PHC setting depending on context [1]. For example, the community's need, the type of model or availability of PHC services will be different in developed settings as compared to LMIC.

Risk of bias in individual studies

For this review, we drew on the use of Tong et al.'s criteria for reporting of qualitative studies [19], on Grant et al.'s proposed framework of minimal requirements for the reporting of process evaluations of cluster randomised controlled trials [17] and on MRC recommendations for process evaluations of complex interventions [4]. Combining insights from these papers, a form of appraisal for risk of bias was derived (see Additional file 5). For the purposes of this review in examining the use of process evaluations alongside RCTs in PHC, studies were not excluded based on quality [20]. Instead, the quality of the studies is presented as a risk of bias graph (low, unclear and high risk) [41].

Data synthesis

This will involve the aggregation or synthesis of qualitative findings to generate a set of statements that represent that aggregation and categorisation of these findings on the basis of similarity in meaning and contexts. These categories will then be subjected to thematic synthesis in order to produce a single comprehensive set of synthesised findings that can be used as a basis for evidence-based practice. The synthesis of these qualitative data aims to satisfy the criteria established for the reporting of the synthesis of qualitative health research [18]. Abstracted quantitative data (e.g. number of positive trials) will be presented together with a descriptive narrative form including tables and figures to aid in data presentation where appropriate. We will examine how authors address potential bias through a narrative synthesis how well these are reported in the papers and strategies that may have been employed to mitigate this (e.g. triangulation of key findings). Depending on papers included, there may be subgroup analysis of further exploration of any differences of the barriers and facilitators to intervention implementation by context such as indigenous versus non-indigenous and of developed settings as compared to LMIC.

Discussion

There is a global call for PHC reform in the areas of service delivery, public policy and leadership to enable greater equity and improved health to different populations. To effect this change will require complex interventions involving multiple players (clinicians, community, allied health professionals, policy makers), disciplines (e.g. education, health) and what is successful in one context may not be suitable in another. Process evaluations

conducted alongside RCTs of complex health interventions are valuable in determining whether a complex intervention should be scaled up or modified for other contexts.

The conduct of process evaluations is still a dynamic area with no clear defined method, partly due to the spectrum of methods (e.g. observation, interviews and routine monitoring data). De Silva et al. in 2014 outlined the integration of the Theory of Change into the MRC framework for complex interventions, and one of its aims was to combine '*process and effectiveness indicators into a single analysis which can help untangle whether, how and why an intervention has an impact in a particular context, and whether it may be suitable for scale up or adaptation for new settings*' [42]. Moreover, in regard to future scale up of complex interventions, economic issues pertinent to stakeholders (e.g. patients and providers) would be crucial to policy makers and funders—while this has not been traditionally incorporated together with process evaluations, it would be helpful to see if it has been done [35, 43].

Process evaluations of complex interventions have been increasing in recent years and seem to be variable in objectives, methodology and quality. The MRC guidance in the conduct of process evaluations and in the interpretation of the RCT outcomes may be helpful for researchers to aid in the implementation of effective interventions beyond the research setting. This protocol outlines our methods and design in our efforts to systematically consolidate the collective experience of researchers in this field in conducting, analysing and reporting process evaluations by assembling the findings within the MRC's process evaluation recommendations and to understand previous challenges and potential solutions in the implementation of evidence-based complex interventions in PHC according to context.

Additional files

Additional file 1: PRISMA-P checklist. (DOC 82 kb)

Additional file 2: Example of search strategy. (PDF 314 kb)

Additional file 3: Eligibility forms. (DOCX 14 kb)

Additional file 4: Data extraction form comprising of four tables. (DOC 34 kb)

Additional file 5: Form of appraisal for risk of bias. (DOC 33 kb)

Abbreviations

LMIC, low- and middle-income countries; MRC, Medical Research Council UK; PHC, primary health care; RCT, randomised controlled trial

Acknowledgements

Not applicable.

Funding

This systematic review forms part of HL's PhD thesis and is not externally funded or commissioned.

Availability of data and materials

Not applicable.

Authors' contributions

HL conceived the study, conducted the scoping searches, designed and piloted the forms and drafted the manuscript; she manages the overall study and will be involved in the study selection, data extraction, synthesis and analysis. JM assisted in the scoping searches, piloted the data and quality appraisal forms, contributed to the drafts of the manuscript and is involved in the study selection. MH provided guidance to HL in the overall design of the study, assisted in refining the data extraction forms and drafted the manuscript. AH helped revise the manuscript and will contribute to the study selection, data extraction and synthesis. TL contributed to the early drafts of the manuscript, revised the manuscript and will contribute to the study selection. DP contributed to the early drafts of the manuscript and revised the manuscript. SJ conceived the study, provided oversight to HL and drafted the manuscript. All authors read and approved the manuscript.

Authors' information

HL had been funded by a University of Sydney Postgraduate scholarship and is currently funded by a National Health and Medical Research Council (NHMRC) scholarship. JM is a recipient of a PhD scholarship from The Australian Prevention Partnership Centre. MH is a recipient of a National Heart Foundation Future Leader Fellowship, Level 2 (100034, 2014–2017). SJ is the recipient of an NHMRC Senior Research Fellowship. TL is the recipient of a NHMRC fellowship. DP is the recipient of a Harkness Fellowship.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Received: 29 April 2016 Accepted: 3 August 2016
Published online: 15 August 2016

References

1. World Health Organisation. The World Health Report—research for universal health coverage. 2013. available from: http://www.who.int/whr/2013/report/en/. Accessed 19 Dec 2014.
2. Chalkidou K, Tunis S, Whicher D, Fowler R, Zwarenstein M. The role for pragmatic randomized controlled trials (pRCTs) in comparative effectiveness research. Clin Trials. 2012;9(4):436–46.
3. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M, et al. Developing and evaluating complex interventions: the new Medical Research Council guidance. BMJ. 2008;337:a1655. Pubmed Central PMCID: 2769032.
4. Moore GF, Audrey S, Barker M, Bond L, Bonell C, Hardeman W, et al. Process evaluation of complex interventions: Medical Research Council guidance. BMJ. 2015;350:h1258. Pubmed Central PMCID: 4366184.
5. Bonell C, Oakley A, Hargreaves J, Strange V, Rees R. Assessment of generalisability in trials of health interventions: suggested framework and systematic review. BMJ. 2006;333(7563):346–9. Pubmed Central PMCID: 1539056, Epub 2006/08/12. eng.
6. Liu H, Massi L, Laba TL, Peiris D, Usherwood T, Patel A, et al. Patients' and providers' perspectives of a polypill strategy to improve cardiovascular prevention in Australian primary health care: a qualitative study set within a pragmatic randomized, controlled trial. Circ Cardiovasc Qual Outcomes. 2015;8(3):301–8.
7. Tugwell P, Robinson V, Grimshaw J, Santesso N. Systematic reviews and knowledge translation. Bull World Health Organ. 2006;84(8):643–51. Pubmed Central PMCID: 2627444.
8. Tetroe JM, Graham ID, Foy R, Robinson N, Eccles MP, Wensing M, et al. Health research funding agencies' support and promotion of knowledge translation: an international study. Milbank Q. 2008;86(1):125–55. Pubmed Central PMCID: 2690338.
9. Cordero C, Delino R, Jeyaseelan L, Lansang MA, Lozano JM, Kumar S, et al. Funding agencies in low- and middle-income countries: support for knowledge translation. Bull World Health Organ. 2008;86(7):524–34. Pubmed Central PMCID: 2647493.

10. Evans DB, Edejer TT, Adam T, Lim SS. Methods to assess the costs and health effects of interventions for improving health in developing countries. BMJ. 2005;331(7525):1137–40. Pubmed Central PMCID: 1283282.
11. Moore G AS, Barker M, Bond L, Bonell C, Hardeman W, Moore L, O'Cathain A, Tinati T, Wight D, Baird J. Process evaluation of complex interventions: Medical Research Council guidance. MRC Population Health Science Research Network, London, 2014. Available on: https://www.mrc.ac.uk/documents/pdf/mrc-phsrn-process-evaluation-guidance-final/. Accessed 22 Jan 2015.
12. Lewin S, Glenton C, Oxman AD. Use of qualitative methods alongside randomised controlled trials of complex healthcare interventions: methodological study. BMJ. 2009;339:b3496. Pubmed Central PMCID: 2741564.
13. Curry LA, Nembhard IM, Bradley EH. Qualitative and mixed methods provide unique contributions to outcomes research. Circulation. 2009;119(10):1442–52.
14. Snowdon C. Qualitative and mixed methods research in trials. Trials. 2015; 16:558. Pubmed Central PMCID: 4672490.
15. O'Cathain A, Goode J, Drabble SJ, Thomas KJ, Rudolph A, Hewison J. Getting added value from using qualitative research with randomized controlled trials: a qualitative interview study. Trials. 2014;15:215. Pubmed Central PMCID: 4059032.
16. Oakley A, Strange V, Bonell C, Allen E, Stephenson J, Team RS. Process evaluation in randomised controlled trials of complex interventions. BMJ. 2006;332(7538):413–6. Pubmed Central PMCID: 1370978.
17. Grant A, Treweek S, Dreischulte T, Foy R, Guthrie B. Process evaluations for cluster-randomised trials of complex interventions: a proposed framework for design and reporting. Trials. 2013;14(1):15. Pubmed Central PMCID: 3600672. Epub 2013/01/15. eng.
18. Tong A, Flemming K, McInnes E, Oliver S, Craig J. Enhancing transparency in reporting the synthesis of qualitative research: ENTREQ. BMC Med Res Methodol. 2012;12:181. Pubmed Central PMCID: 3552766.
19. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care. 2007;19(6):349–57.
20. Petticrew M. Time to rethink the systematic review catechism? Moving from 'what works' to 'what happens'. Syst Rev. 2015;4:36.
21. Kirby T. Profile: Australia's George Institute for Global Health. Lancet. 2015;385(9977):1498.
22. Alim M, Lindley R, Felix C, Gandhi DB, Verma SJ, Tugnawat DK, et al. Family-led rehabilitation after stroke in India: the ATTEND trial, study protocol for a randomized controlled trial. Trials. 2016;17(1):13. Pubmed Central PMCID: 4704425.
23. Praveen D, Patel A, Raghu A, Clifford GD, Maulik PK, Mohammad Abdul A, et al. SMARTHealth India: development and field evaluation of a mobile clinical decision support system for cardiovascular diseases in rural India. JMIR MHealth UHealth. 2014;2(4):e54. Pubmed Central PMCID: 4275493.
24. Laba TL, Hayes A, Lo S, Peiris DP, Usherwood T, Hillis GS, et al. An economic case for a cardiovascular polypill? A cost analysis of the Kanyini GAP trial. Med J Aust. 2014;201(11):671–3.
25. Lawoyin TO, Lawoyin OO. Translation of research into reality in sub-Saharan Africa. Lancet. 2013;381(9884):2146–7.
26. Peiris D, Usherwood T, Panaretto K, Harris M, Hunt J, Redfern J, et al. Effect of a computer-guided, quality improvement program for cardiovascular disease risk management in primary health care: the treatment of cardiovascular risk using electronic decision support cluster-randomized trial. Circ Cardiovasc Qual Outcomes. 2015;8(1):87–95.
27. Neubeck L, Coorey G, Peiris D, Mulley J, Heeley E, Hersch F, et al. Development of an integrated e-health tool for people with, or at high risk of, cardiovascular disease: The Consumer Navigation of Electronic Cardiovascular Tools (CONNECT) web application. International journal of medical informatics. 2016 [Epub ahead of print]
28. Chow CK, Redfern J, Hillis GS, Thakkar J, Santo K, Hackett ML, et al. Effect of lifestyle-focused text messaging on risk factor modification in patients with coronary heart disease: a randomized clinical trial. JAMA. 2015;314(12):1255–63.
29. World Health Organisation. Global status of non-communicable diseases World Health Organisation, 2014. Available on: http://www.who.int/nmh/publications/ncd-status-report-2014/en/. Accessed 25 May 2016.
30. Bero LA, Grilli R, Grimshaw JM, Harvey E, Oxman AD, Thomson MA. Closing the gap between research and practice: an overview of systematic reviews of interventions to promote the implementation of research findings. The Cochrane Effective Practice and Organization of Care Review Group. BMJ. 1998; 317(7156):465–8. Pubmed Central PMCID: 1113716. Epub 1998/08/14. eng.
31. Best A, Greenhalgh T, Lewis S, Saul JE, Carroll S, Bitz J. Large-system transformation in health care: a realist review. Milbank Q. 2012;90(3):421–56. Pubmed Central PMCID: 3479379.
32. Christopher Dye TB, David Evans, Anthony Harries, Christian Lienhardt, Joanne McManus, Tikki Pang, Robert Terry, Rony Zachariah. The world health report 2013: research for universal health coverage. World Health Organisation; 2013. http://www.who.int/whr/2013/report/en/.
33. Gelijns AC, Gabriel SE. Looking beyond translation—integrating clinical research with medical practice. N Engl J Med. 2012;366(18):1659–61.
34. Shamseer L, Moher D, Clarke M, Ghersi D, Liberati A, Petticrew M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. BMJ. 2015;349:g7647.
35. Anderson LM, Oliver SR, Michie S, Rehfuess E, Noyes J, Shemilt I. Investigating complexity in systematic reviews of interventions by using a spectrum of methods. J Clin Epidemiol. 2013;66(11):1223–9.
36. World Health Organisation. The Declaration of Alma-Ata. 1978. Available from: http://www.who.int/publications/almaata_declaration_en.pdf. Accessed 16 Apr 2015.
37. World Health Organisation. Primary health care [16 april 2015]. Available from: http://www.who.int/topics/primary_health_care/en/. Accessed 25 May 2016.
38. Davy C, Bleasel J, Liu H, Tchan M, Ponniah S, Brown A. Effectiveness of chronic care models: opportunities for improving healthcare practice and health outcomes: a systematic review. BMC Health Serv Res. 2015;15:194. Pubmed Central PMCID: 4448852.
39. Davies P, Walker AE, Grimshaw JM. A systematic review of the use of theory in the design of guideline dissemination and implementation strategies and interpretation of the results of rigorous evaluations. Implement Sci. 2010;5: 14. Pubmed Central PMCID: 2832624, Epub 2010/02/26. eng.
40. Petticrew M, Anderson L, Elder R, Grimshaw J, Hopkins D, Hahn R, et al. Complex interventions and their implications for systematic reviews: a pragmatic approach. Int J Nurs Stud. 2015;52:1211–6.
41. Katikireddi SV, Egan M, Petticrew M. How do systematic reviews incorporate risk of bias assessments into the synthesis of evidence? A methodological study. J Epidemiol Community Health. 2015;69(2):189–95. Pubmed Central PMCID: 4316857.
42. De Silva MJ, Breuer E, Lee L, Asher L, Chowdhary N, Lund C, et al. Theory of change: a theory-driven approach to enhance the Medical Research Council's framework for complex interventions. Trials. 2014;15:267. Pubmed Central PMCID: 4227087.
43. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implement Sci. 2009;4:50. Pubmed Central PMCID: 2736161.

NOTES

Thank you