Adjusting for Selection Effects in Epidemiologic Studies

Why Sensitivity Analysis is the Only "Solution"

Sara Geneletti,[a] *Alexina Mason,*[b] *and Nicky Best*[b]

As participation rates decline, epidemiologists are faced with a growing challenge in interpreting the data that are available. Barnighausen et al[1] and Chaix et al[2] provide thoughtful case studies in which the implications of survey nonparticipation are carefully considered and statistical models chosen to adjust for likely bias. Will papers such as these help to persuade epidemiologists, on a routine basis, to pay more than lip service to issues of selection? The impact of selection bias may often be quite weak and the adjustment methods technically difficult. However, it is essential for researchers to think formally about the possible sources of bias in the data they plan to analyze and to assess the sensitivity of their conclusions to these potential biases.

The 2 papers illustrate the use of different variants of selection models, which is just one of a number of approaches open to epidemiologists for adjusting for possible bias. But, practically speaking, does it matter which adjustment method is used? Is some sort of adjustment better than none? Certainly, as nonparticipation increases, so do the risks that an analysis based only on complete cases will result in biased inference and invalid conclusions. Thus, some form of adjustment should be considered. The choice of adjustment method depends on plausible assumptions regarding the nature of the nonparticipation, and on the type of additional sources of data that are available. However, any chosen model will generally be based on untestable assumptions, because by definition we do not observe the characteristics of nonparticipants. For this reason, any method that attempts to correct for nonparticipation bias is essentially a sensitivity analysis. It is perfectly possible that a different set of assumptions about the selection process will lead to different adjustments of the parameters of interest, and the implications of this possibility should always be explored and reported.

IDENTIFYING POTENTIAL SOURCES OF BIAS RESULTING FROM NONPARTICIPATION

In both papers,[1,2] the researchers thought first about the structural assumptions they had to make about nonparticipation, and second about what data they could use to inform a participation model. Only then could they develop a procedure to adjust for nonparticipation bias. The structural assumptions refer to the mechanism that introduces bias: Are the participants systematically different from the nonparticipants on the variables of substantive interest? If so, how does this difference manifest itself? Graphical models, such as directed acyclic graphs (DAGs), can be a useful tool for exploring these issues, and indeed Chaix et al[2] use them to identify "collider bias." We return to the use of such DAGs below.

From the [a]Department of Statistics, London School of Economics and Political Science, London, United Kingdom; and [b]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College, London, United Kingdom.

Supported by ESRC: RES-576–25–0015, M (to A.M. and N.B.).

Correspondence: Sara Geneletti, Department of Statistics, London School of Economics and Political Science, Houghton Street, WC2A 2A4, London, United Kingdom. E-mail: s.geneletti@lse.ac.uk.

ISSN: 1044-3983/11/2201-0036
DOI: 10.1097/EDE.0b013e3182003276

TYPES OF ADDITIONAL DATA

Information about nonparticipation can be thought of as coming in 2 types, exemplified in the 2 papers[1,2]—internal and external. Internal information comprises data that are available on all the individuals who are eligible to participate, regardless of whether they provide any information relating to the substantive question. Typically, this situation occurs when the study is conducted within a cohort (eg, a nested case–control study) or a census, or when individuals in previous sweeps of a longitudinal study drop out. In this case, we have some individual-level information about the nonparticipants that might be relevant to their nonparticipation. In the paper on HIV,[1] additional available data included numbers living in a household and interviewer identity, both of which were used to inform the selection model.

There are also situations—such as cross-sectional health surveys, cohort studies, or case–control studies set outside of cohorts—where no individual-level information on nonparticipants is available. Fortunately, due to the large amount of data routinely collected in public health, it is often possible to find data that cover the same population as that of the study under investigation. This is external information, which comes from a different data source and does not include information on the individuals themselves, but may be of use for modeling nonparticipation. In fact, it is often worth thinking about this aspect during study design, and collecting information with a particular auxiliary data source in mind, in such a way that the study can be linked to these data sources in the analysis phase. This set-up is described in the paper on neighborhood effects by Chaix et al,[2] in which individuals were recruited without a definite sampling frame, and a census provides external information based on neighborhood of residence of eligible participants.

GRAPHICAL MODELS CAN HELP IDENTIFY MECHANISMS LEADING TO BIAS

DAGs are a useful tool for visualizing complex relationships among variables and for understanding potential sources of bias. There are a number of papers that can be used as recipes to identify what variables are likely to cause bias in a dataset.[3,4] Recent work by Hernán et al[4] describes very clearly how to determine whether a study is likely to be suffering from nonparticipation bias. When this is the case, the variable that indicates participation is a "collider." In the papers by Chaix et al[2] and Barnighausen et al,[1] the DAG that describes the relationships among the variables of interest includes a collider, indicating that selection bias is a potential problem, as we illustrate below.

Figures 1 and 2 represent the relationships between the variables involved in the problems in the papers by Barnighausen et al[1] and Chaix et al,[2] respectively. Figure 1A and B mirror Figure 2A and B, showing how participation bias manifests itself in the same way in both papers. In both cases, X and U are the observed and unobserved variables, respectively; S is the selection indicator; and Y the outcome of interest (HIV or diabetes status). In the analysis by Barnighausen et al, under the Heckman model, U can be understood as the unknown correlation between the selection and observed variables, whereas in the model by Chaix et al, U represents the unobserved neighborhood effects.

Figures 1A and 2A show both observed and unobserved variables. Figures 1B and 2B, however, show only the observed variables and the implied dependence due to not conditioning on unobserved variables. The latter DAGs demonstrate the potential for selection bias, as S is a collider between the outcome Y and the observed covariates X.

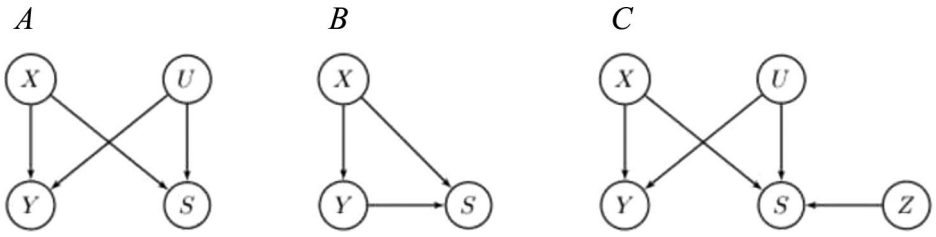

FIGURE 1. DAG representing the analysis by Barnighausen et al.[1] X are the observed characteristics of the respondents and U is the unobserved correlation. U can also be viewed as unobserved characteristics. S is the selection indicator and Y is the HIV status. Z represents the selection variables, interviewer identity, or identity of an interviewer of a member of the household.

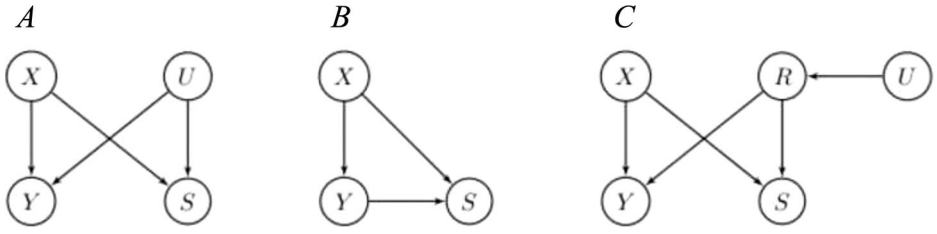

FIGURE 2. DAG representing the analysis by Chaix et al.[2] X are the observed neighborhood effects and U are the unobserved neighborhood effects. S is the selection indicator and Y is diabetes status. R represents the random effects.

Figures 1C and 2C represent the 2 approaches used to tackle participation bias. By introducing selection variables Z in Figure 1C such that the Heckman assumption of independence of Z and Y holds, Barnighausen et al[1] are able to identify and estimate the unobserved correlation and adjust for selection bias. Chaix et al[2] chose a different approach to adjusting for the bias, as shown in Figure 2C, by finding a proxy for the unobserved neighborhood effects in the form of the random effects R.

SELECTION OF APPROPRIATE MODELING METHOD

Only when the reasons for, and implications of, the nonparticipation have been thought through thoroughly, is the analyst in a position to select an appropriate modelling method. The choice depends on whether the resulting missingness can plausibly be assumed to be missing at random[5] (ie, the probability of being missing is not dependent on unobserved data, given the observed data). For example, in the paper by Barnighausen et al,[1] missing at random means that the unobserved correlation is 0 and U disappears from the DAG in Figure 1A. In this case, there is often no need to model the participation process, and options include multiple imputation,[6] reweighting procedures such as inverse probability weighting[7] or poststratification,[8] and bias-modeling techniques.[9]

Barnighausen et al[1] considered that the missing HIV data from the nonresponders was likely to be missing not at random[5] (ie, the probability of being missing is dependent on unobserved data, given the observed data), and so a method that allowed the joint modeling of the participation process and the substantive question was required. Chaix et al[2] also favored this joint-model approach, as the neighborhood random effects were thought to influence both their study participation model and their diabetes model. As we have discussed, both groups of researchers use a selection model but with different forms, illustrating how the modeling choice is problem-specific, as well as dependent on assumptions made and the type of additional data available. A third option for modeling nonresponse that is missing not at random is to explicitly model the link between Y and S in Figures 1B and 2B, by including Y as a predictor in the selection equation.[10]

Selection models can be implemented within traditional (Barnighausen et al[1]) or Bayesian (Chaix et al[1]) estimation frameworks. A Bayesian approach provides the option of incorporating information through expert priors, which can be formed through elicitation or literature search. For instance, in the HIV paper, data from the Malawi study on the probability of refusing an HIV test given HIV status could be incorporated into an informative prior on the covariance matrix of the Heckman model.

SENSITIVITY ANALYSIS

As we have stressed, model choice and hence results are dependent on the assumptions made. Unfortunately, it is not possible to test whether missing data is missing at random or not at random—despite the slightly misleading impression given by the tests carried out by Barnighausen et al[1]—because identification of the correlation between HIV status and participation is completely dependent on the choice of Z variable (exclusion restriction) and the distributional assumptions of the substantive and selection models. Consequently, it is essential that the robustness of results is tested by fitting a range of models that incorporate varying assumptions. This can be as simple as the initial analyses of the HIV data,[1] where estimates were calculated assuming either that the missing individuals were all HIV-positive or all HIV-negative. A more sophisticated approach would, for example, involve varying the form of the different parts of a joint model. We have found that a Bayesian approach is very conducive to these types of complex analysis in that the modular setup allows various assumptions about the nonparticipation model or the analysis model to be explored relatively easily. Our experience suggests that varying the functional form of either the analysis or participation model can substantially alter results (A Mason, S Richardson, I Plewis, et al, unpublished data). In the analysis by Barnighausen et al,[1] which uses the frequentist framework, it would be interesting to explore the implications of using different exclusion variables.

CONCLUSIONS

With increasing rates of nonparticipation in surveys and studies, it becomes more important that epidemiologists recognize the inherent uncertainty and potential for bias that accompany nonresponse. A mindset that bases conclusions on a single "best" model needs to be replaced by one that presents a range of models encompassing different plausible assumptions, or equivalently a "base model" accompanied by a series of sensitivity analyses. It may turn out that all the results are robust to a range of assumptions, but unfortunately there is no way of knowing this before carrying out the extended analysis. The challenge for the researcher is to choose the most appropriate statistical tool or approach for their particular problem, given their subject knowledge, and utilizing as much available additional information as possible. Epidemiologists would be more likely to go down this route if more practical advice and real examples showing its value are available, and the 2 papers discussed here contribute to this process. Equally important is access to, and understanding of, software that allows the plausibility of different assumptions about nonparticipation to be explored.

Chaix et al[2] and Barnighausen et al[1] each conclude that their method should be routinely used. We contend that the specific method is not important (although it should be

appropriate to the situation), but that routine practice should follow the key principles of thinking about the selection process and assessing sensitivity to different assumptions. To quote the advice of Allen and Holland[11] given to educational researchers over 20 years ago: "You must be prepared to think as hard about your nonrespondents as you do about your substantive research and to incorporate this into a sensitivity analysis. Otherwise, you have not handled selection bias but have only ignored it."

ABOUT THE AUTHORS

SARA GENELETTI is a lecturer in Statistics at the London School of Economics. Her research focuses on bias adjustment for causal inferences using DAGs and evidence synthesis. ALEXINA MASON is a research associate on the BIAS project, and her research interests include Bayesian methods for modeling missing data mechanisms. NICKY BEST is Professor of Statistics and Epidemiology at Imperial College London. She is Director of the BIAS project (www.bias-project.org.uk) which is developing statistical methodology to model biases in observational data.

REFERENCES

1. Barnighausen T, Bor J, Wandira-Kazibwe S, Canning D. Correcting HIV prevalence estimates for survey nonparticipation: using Heckman-type selection models. *Epidemiology*. 2011;22:27–35.
2. Chaix B, Billaudeau N, Thomas F, et al. Neighborhood effects on type 2 diabetes: correcting bias from neighborhood effects on participation. *Epidemiology*. 2011;22:18–26.
3. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology*. 1999;10:37–48.
4. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology*. 2004;15:615–625.
5. Rubin DB. Inference and missing data. *Biometrika*. 1976;63:581–592.
6. Kenward MG, Carpenter J. Multiple imputation: current perspectives. *Stat Methods Med Res*. 2007;16:199–218.
7. Robins JM, Finkelstein D. Correcting for non-compliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics*. 2000;56:779–788.
8. Geneletti S, Richardson S, Best N. Adjusting for selection bias in retrospective, case–control studies. *Biostatistics*. 2009;10:17–31.
9. Turner RM, Spiegelhalter DJ, Smith GC, Thompson SG. Bias modelling in evidence synthesis. *J R Stat Soc Ser A Stat Soc*. 2009;172:21–47.
10. Daniels MJ, Hogan JW. *Missing Data in Longitudinal Studies Strategies for Bayesian Modeling and Sensitivity Analysis*. Boca Raton, FL: Chapman & Hall 2008;167–181.
11. Allen NL, Holland PW. Exposing our ignorance: the only "solution" to selection bias. *J Educ Stat*. 1989;14:141–145.