Sara Geneletti, Alexina Mason and Nicky Best

Adjusting for selection effects in epidemiologic studies: why sensitivity analysis is the only "solution"

Article (Accepted version)
(Refereed)

Original citation:
Geneletti, Sara and Mason, Alexina and Best, Nicky (2011) *Adjusting for selection effects in epidemiologic studies: why sensitivity analysis is the only "solution".* *Epidemiology*, 22 (1). pp. 36-39. ISSN 1044-3983
DOI: 10.1097/EDE.0b013e3182003276

Introduction

Chaix et al[1] and Barnighausen et al[2] provide thoughtful case studies in which the implications of survey non-participation are carefully considered and statistical models chosen to provide adjustment for likely bias. But will papers such as these help to persuade epidemiologists to pay more than lip service to the issues of selection on a routine basis? The impact of selection bias may often be quite weak and the adjustment methods may be technically difficult. However we argue that it is essential for researchers to formally think about the possible sources of bias in the data they plan to analyse and to assess sensitivity of their conclusions to these potential biases.

The two 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 the adjustment method used matter? Is some sort of adjustment better than none? Certainly, as non-participation increases, so do the risks that an analysis based only on complete cases will result in biased inference and invalid conclusions, and so some form of adjustment should be considered. The choice of adjustment method depends on the assumptions that are considered plausible regarding the nature of the non-participation and 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 primary interest of the non-participants. Thus any method that attempts to correct for non-participation 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 should always be explored and reported.

Identifying potential sources of bias resulting from non-participation

In both papers[1,2] the researchers thought first about the structural assumptions they had to make about the non-participation, and second about what data they could use to inform a participation

model before developing a procedure to adjust for non-participation bias. The structural assumptions refer to the mechanism that introduces bias, i.e. we must seek to answer the questions: Are the participants systematically different from the non-participants on the variables of substantive interest? If so, how does this difference manifest itself? We have found that graphical models, such as directed acyclic graphs (DAGs), are a useful tool for exploring these issues, and indeed Chaix et al use them to identify "collider bias". We discuss the use of such DAGs later below.

Types of additional data

Information about non-participation can be thought of as coming in two types which are exemplified in the two papers:[1,2] internal and external. Internal information comprises data which is available on all the individuals who are eligible to participate in a study, regardless of whether they provide any information relating to the substantive question. Typically this situation occurs when the study is conducted within a cohort (e.g. 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 non-participants which might be relevant to their non-participation. In the HIV paper[2], 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, for example, cross-sectional health surveys, cohort studies or case-control studies that are set outside of cohorts, where no individual level information on the non-participants is available. Fortunately, due to the large amount of data that are routinely collected in public health, it is often possible to find data that covers 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 modelling non-participation. In fact it is often worth thinking about this aspect during the study design, and to collect information with a particular auxiliary data source in mind, in such a way that linking

the study to these data sources is easy in the analysis phase. This set-up is described in the paper on neighbourhood effects by Chaix et al[1] where individuals are recruited without a definite sampling frame, and a census provides external information based on neighbourhood of residence of eligible participants.

Graphical models can help identify mechanisms leading to bias

DAGs are becoming increasingly popular in the epidemiologic literature. They are very useful for visualizing complex relationships between variables and for understanding potential sources of bias. There now exists a number of papers that can be used as recipes to identify what variables are likely to cause bias in a data-set.[3,4] Recent work by Hernan et al[4] describes very clearly how to determine whether a study is likely to be suffering from non-participation bias. When this is the case, the variable that indicates participation is a "collider". In both Chaix et al and Barnighausen et al, the DAG that describes the relationships between the variables of interest has participation as a collider, indicating that selection bias is a potential problem, as we illustrate below.

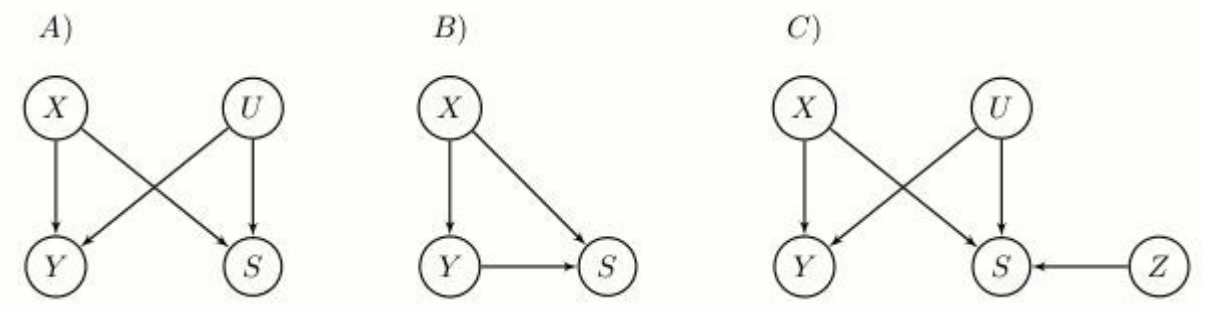

Figure 1: DAG representing the situation in Barnighausen et al. X are the observed characteristics of the respondents and U is the unobserved correlation. U can also be viewed as unobserved characteristics. S is the selection indicator and Y is the HIV status. Z are the selection variables, interviewer identify or identify of an interviewer of a member of the household.

Figures 1 and 2 represent the relationships between the variables involved in the problems in the papers by Barnighausen et al and Chaix et al respectively. Figures 1a, 1b and 2a, 2b mirror one

another and show how participation bias manifests itelf in the same way in both papers. In particular 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 Barnighausen et al, under the Heckman model, U can be understood as the unknown correlation between the selection and observed variables, whereas in Chaix et al, U are the unobserved neighbourhood 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.

Figures 1c and 2c differ because they represent the two 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 are able to identify and estimate the unobserved correlation and adjust for selection bias. Chaix et al choose a different approach to adjusting for the bias in Figure 2c by finding a proxy for the unobserved neigbourhood effects in the form of the random effects R.

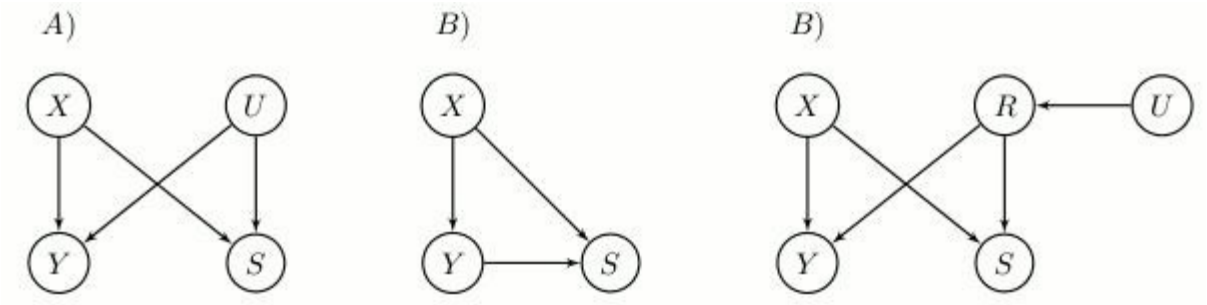

Figure 2: DAG representing the situation in Chaix et al. X are the observed neigbourhood effects and U are the unobserved neighbourhood effects. S is the selection indicator and Y is diabetes status. R are the random effects.

Selection of appropriate modelling method

Only when the reasons for, and implications of, the non-participation 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, MAR[5] (i.e. the probability of being missing is not dependent on unobserved data, given the observed data). For example, in Barnighausen et al, MAR 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], re-weighting procedures such as inverse probability weighting[7] or post-stratification[8] and bias modelling techniques[9].

Barnighausen et al considered that the missing HIV data from the non-responders was likely to be missing not at random, MNAR[5] (i.e. the probability of being missing is dependent on unobserved data, given the observed data), so a method which allowed the joint modelling of the participation process and the substantive question was required. Chaix et al also favoured this joint model approach, as the neighbourhood random effects were thought to influence both their study participation model and their diabetes model. As we have discussed, both use a selection model, but the form differs, illustrating how the modelling choice is problem specific and dependent on assumptions made and the type of additional data available. A third option for modelling MNAR non-response 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) or Bayesian (Chaix et al) 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 MAR or MNAR (despite the slightly misleading impression given by the tests carried out by Barnighausen et al, since 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 which incorporate varying assumptions. This can be as simple as the initial analyses of the HIV data[2], where estimates were calculated assuming either that the missing individuals were all HIV positive or all HIV negative, or can be sophisticated and, 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, as the modular setup allows different assumptions about the non-participation 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 and N Best, Strategy for modelling non-random missing data mechanisms in observational studies using Bayesian methods, *working paper, 2010)*. In Barnighausen et al, which uses the frequentist framework, it would be interesting to explore the implications of using different exclusion variables.

Conclusions

With increasing rates of non-participation in surveys and studies, it becomes more important that epidemiologists recognise the inherent uncertainty and potential for bias that accompanies non-response. A mindset that bases conclusions on a single 'best' model needs to change to 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 different assumptions, but unfortunately there is no way of knowing this without

carrying out the extended analysis. The challenge for the researcher is to choose the most appropriate statistical tool/approach for their particular problem, given their subject knowledge, utilising as much available additional information as possible. Epidemiologists are more likely to go down this route if more practical advice and real examples which show its value are available, and the two papers discussed here will contribute to this process. Equally important is access to, and understanding of, software that allows the plausibility of different assumptions about non-participation to be explored.

Chaix et al and Barnighausen et al each conclude that *their* method should be routinely used. We contend that the specific method is not so important, although it should be appropriate, 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 non-respondents 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."

References

1. Chaix B, Billaudeau N, Thomas F, et al. A joint modeling of neighborhood effects on participation in the RECORD Cohort Study and neighborhood effects on type 2 diabetes: bias assessment and correction. *Epidemiology*.

2. Barnighausen T, Bor J, Wandira-Kazibwe S, Canning D. Correcting HIV prevalence estimates for survey non-participation: an application of Heckman-type selection models to the Zambian Demographic and Health Survey. *Epidemiology.*

3. Greenland S, Pearl J, Robins JM. Causal Diagrams for Epidemiologic Research. *Epidemiology*. 1999; **10(1)**:37-48.

4. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology*. 2004; **15**(5):615-25.

5. Rubin DB. Inference and Missing Data. *Biometrika*. 1976; 63(3): 581-92.

6. Kenward MG and Carpenter J, Multiple imputation: current perspectives, *Statistical Methods in Medical Research*. 2007; **16**(3):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**(3):779-88.

8. Geneletti S, Richardson S, Best N. Adjusting for selection bias in retrospective, case–control studies. *Biostatistics*. 2009; **10**(1): 17-31.

9. Turner RM, Spiegelhalter DJ, Smith GCS and Thompson SG. Bias modelling in evidence synthesis. *JRSSA*. 2009; **172**(1):21-47

10. Daniels MJ, Hogan JW. Missing Data In Longitudinal Studies Strategies for Bayesian Modeling and Sensitivity Analysis. Chapman & Hall; 2008: 167-181.

11. Allen NL and Holland PW. Exposing our ignorance: the only "solution" to selection bias. *Journal of Educational Statistics*. 1989; **14**(2):141-45