



Education Corner

Reflection on modern methods: selection bias—a review of recent developments

Claire Infante-Rivard^{1*} and Alexandre Cusson²

¹Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, QC, Canada and ²Research Centre, Centre Hospitalier Universitaire (CHU) Sainte-Justine, Montréal, QC, Canada

*Corresponding author. Department of Epidemiology, Biostatistics and Occupational Health, Faculty of Medicine, McGill University, 1020 Avenue Des Pins Ouest, Montréal, QC H3A1A2, Canada. E-mail: claire.infante-rivard@mcgill.ca

Editorial decision 31 May 2018; Accepted 11 June 2018

Abstract

Selection bias remains a more difficult bias to understand than confounding or measurement error. Past definitions have not always been illuminating and a simple method (such as the change-in-estimate method for confounding) has not been available to determine its presence and magnitude in the study sample. A better understanding of the nature of the bias has led to the definition of endogenous selection bias. It is the result of conditioning on a collider variable, itself caused by two other variables; the latter variables become spuriously associated. Conditioning on a variable in the analysis that is a collider or on an indicator of sample selection has the same effect. Note that selection bias is possible even in the absence of a collider, but in the presence of endogenous selection bias, the concern is whether it is possible to identify a causal effect in the sample. Conditions have been outlined to determine that. However, even if conditions are met to identify a causal effect in the study sample, its generalization to a defined target population is not a given.

We discuss the concept of endogeneity and the sources of endogenous selection bias in observational studies. We then briefly address the terms generalizability, target population (or alternative formulations) and transportability. We outline the explicit conditions to identify causal effects in studies affected by selection bias: they involve exchangeability between exposed and unexposed and exchangeability between sampled and unsampled. We briefly describe methods to generalize estimated causal effects to the target population. The latter usually require data from the target population. Finally we discuss sensitivity analyses; some are limited to providing an indication of the presence and direction of the bias and others can provide corrected estimates with user-supplied selection bias parameters.

Key words: Selection bias, endogenous selection bias, causal inference, external validity, generalizability, sensitivity analysis

Key Messages

- The definition of selection bias in epidemiology has been inconsistent and is still not as clear as that of confounding. Endogenous selection bias has been proposed: it means conditioning (from adjusting or sample selection) on a common effect of two variables along a path linking exposure and outcome. The common effect is a collider. Although selection bias can be present without a collider, endogenous selection bias covers many common potential sources of this bias arising in studies, from the design to the analysis. As a result the association measure may be biased.
- Conditions (criteria) necessary for the identification of a causal effect (internal validity) in the presence of selection bias, and its generalizability, have been outlined. They include exchangeability on exposure (backdoor criterion) as well as exchangeability between sampled and unsampled (together this is termed the selection backdoor criterion).
- External validity or generalizability of results, from a usually selected sample to a target population (often not defined), has been more of a secondary concern. Nevertheless, results are often generalized as though the study sample was random and the effect causal and, moreover, directly applicable as such to a target population.
- Estimation methods to recover a causal estimate such as the g-formula and inverse probability weighting can be used with a non-random sample, taking into account exchangeability between sampled and unsampled. With information from the target population for the weighting variables in these estimations, the causal effect in the sample can be generalized to a target population.
- With no data available from the target population, sensitivity analyses are still possible to determine the presence and direction of the bias and to obtain corrected estimates.

Introduction

We recognize confounding as a mechanism generating a non-causal path between exposure and outcome. Selection bias can also do that but is more difficult to recognize than confounding or even measurement error. Recently, conceptual features of the bias have been revisited,¹ and conditions under which causal effects can be identified and generalizable from a selected (non-random) sample have been outlined.^{2,3} Finally, methods to recover causal effects have been applied with additional assumptions when used in samples with selection bias; if assumptions are met, generalizing the causal effects from the sample to the target population is possible.^{4,5}

The goals of this paper are first to provide a brief review of recently updated features of selection bias. The review includes: the concept itself; the conditions that will allow identification of causal effects under selection bias; and how familiar estimation methods can be applied in samples with selection bias if a causal effect is identifiable. Finally, generalizing causal effects to a target population as a formal step is also discussed. This latter step requires additional data, usually external to the sample; in the absence of such data, sensitivity analyses to detect the presence of the bias or to obtain corrected estimates with user-supplied parameters are briefly discussed. A summary of the components of this paper is shown in [Box 1](#).

Review of selection bias

Concept

Elwert and Winship¹ suggest the term ‘endogenous selection bias’ as a notion that captures many biases that the epidemiologist usually associates with selection bias, such as sample selection, non-response to questions or non-attendance to a scheduled visit, informative loss to follow-up etc. The term also includes conditioning on a variable in the analysis which is a collider. A collider is the common effect of two variables; commonly, one that is the exposure (or cause of it) and one that is outcome (or cause of it).⁶ Conditioning on a collider that is a variable in the analysis or that is an indicator of selection into study generates similar problems. Examples with sample selection are shown in [Figure 1](#) with directed acyclic graphs (DAG), where *S* is the indicator for sample selection. We return to these examples later.

We first briefly discuss the term endogeneity. Endogeneity occurs when there is a correlation between the error term in a regression and the exposure variable(s). This leads to attributing some of the effect of exposure on outcome to what is in fact due in part to a cause in the error term. The sources of endogeneity are: (i) omitted confounders; (ii) simultaneous equations (where elements on the right-hand side of one equation are also present in the left-hand side of the other equations); (iii) measurement error; and (iv) sample selection as a special case. The latter

Box 1. Sources of selection bias, conditions and estimators to recover causal effects, and some methods for sensitivity analysis

Sources of selection bias from study design to analysis:

- study entry: non-random sample selection in survey, case-control and cohort studies. index-event studies (with diseased subjects looking for a later outcome);
- intermediate level: losses to follow-up from informative censoring (attrition, non-response);
- analysis: conditioning on colliders in propensity score; on mediators that are also colliders; on time-varying confounders that are also colliders.

Conditions to recover causal effect with sample selection: selection backdoor criterion:²

X is the exposure; Y is the outcome; Z is a sufficient set to block all backdoor paths (those that end with an arrow pointing to X); Z is partitioned into Z^+ containing the non-descendants of X; Z^- includes the descendants of X.

To recover the conditional probability of outcome to estimate RR and RD:

- condition 1: Z^+ blocks all backdoor paths from X to Y;
- condition 2: X and Z^+ block all paths between Z^- and Y such that $(Z^- \perp\!\!\!\perp Y|X, Z^+)$ where $\perp\!\!\!\perp$ stands for independent);
- condition 3: X and Z must block all paths between S and Y such that $(Y \perp\!\!\!\perp S|X, Z)$;
- condition 4: Z, X and Y are measured in the data affected by selection bias, and Z is also measured at (or known from) the population level.

To recover the conditional causal odds ratio (COR) in a case-control study:³

- condition 1: a sufficient set Z blocking all backdoor paths between X and Y;
- condition 2: $X \perp\!\!\!\perp S|(Z, Y)$.

Estimation to recover causal effects with sample selection⁴ (see text for formulas):

- g-formula;
- inverse probability of sampling weighted estimator
- inverse probability of censoring weighting
- Heckman sample selection (parametric model for incidental truncation).

Sensitivity analyses with or without user-supplied parameters:

- negative controls (direction and magnitude of bias not measured);
- simple table data analysis using, for example, Stata episens and episensi scripts (deterministic and probabilistic options);
- inverse probability of selection weighting.

can be understood as unmeasured factors (U) influencing both selection into the study (indicator S) and outcome (Y), with exposure under study (X) also influencing S.

Consider the DAG in Figure 1d. Conditioning on collider S has opened a non-causal path linking X and Y such that $X \rightarrow S \leftarrow Z \rightarrow Y$. A non-causal path shows a sequence of arrows linking two variables where not all arrows point away from exposure X and toward the outcome Y. To pursue our explanation of sample selection creating endogeneity, replace Z by an unmeasured U variable such that $X \rightarrow S \leftarrow U \rightarrow Y$. This non-causal path is now open and remains an alternative explanation to the causal path $X \rightarrow Y$ in the study results. The spurious link between X, the exposure, and U, the unmeasured variable, influencing both selection and outcome is a representation of what causes the problem with sample selection. This situation is similar to that of unmeasured confounders as a source of inconsistent exposure estimates.

As we have just seen, on a path between exposure X and outcome Y, conditioning on a collider creates a link between the two variables that have generated the collider leading to a spurious association between them, and opening a non-causal path. Figure 1 shows a number of DAGs describing situations where there is conditioning on sample selection as a collider. We can see that in all of them, because of conditioning on S, non-causal paths are present. If Z, a measured variable, or M and L (in Figure 1e), were substituted for U, an unmeasured variable, the causal effect would not be identified in any of the study typologies represented in the DAGs. We also observe that in Figure 1a and b, the non-causal path $X \rightarrow S \leftarrow Y$ creates an association between X and Y; when none is present (Figure 1a) one is created, and if one is present (Figure 1b) it is likely changed. In studies affected by this pattern of selection bias with no other measured variable, a causal effect cannot be identified.

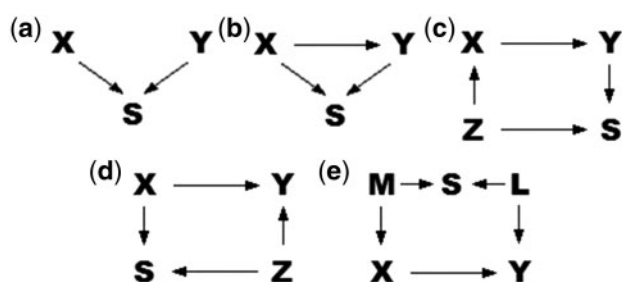


Figure 1. (a)–(e) Directed acyclic graph (DAG). X is the exposure; Y is the outcome; S is an indicator of selection; U is an unmeasured variable; Z, M and L are covariates.

If a causal effect is identifiable as a relative risk or risk difference (conditions discussed later) in a biased sample, such as for example in Figure 1d and e, generalizing it to the target population will usually require additional data. This point is covered later under ‘Conditions to recover causal effect’.

Note that in the collider variable situation (as opposed to collider as an indicator of sample selection), a careful analyst can avoid bias by simply not conditioning on the collider, as this leaves the non-causal path closed. However, with sample selection, one has to conceptualize the problem as similar to a missing data problem.^{7,8} The analysis starts with ‘missing data’ or complete records only; in other words, the sample is already affected with potential selection bias. One has then to determine whether causal effects can be recovered with such a sample.

It is also worth mentioning that whereas conditioning on sample selection as a collider results in endogenous selection bias, the potential for selection bias can also occur with sample selection not being a collider (for example from exposure only; or from a covariate only). Recently, Hernán⁹ explained the situation represented in a DAG shown in Figure 2. X causes Y but does not affect S (an indicator for censoring). An unmeasured variable U affects S and the outcome Y. Here, there is no bias in the sample but the effect cannot be generalized (because selection represented with S cannot be separated from outcome Y; we later describe the explicit criterion for that). The mechanism explaining the findings is an interaction between exposure and the unmeasured variable, resulting in the relationship between S and outcome being different in the censored and uncensored. It is scale-dependent and therefore not represented in the DAG. Cole and Stuart¹⁰ had underscored this mechanism previously when discussing the generalizability of randomized trials.

Returning to the DAGs in Figure 1, for example 1a and 1b, we note that although there are some exceptions¹¹ if both variables pointing to S have an effect in the same direction (positive or negative), then their spurious

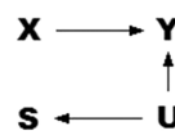


Figure 2. Directed acyclic graph (DAG). X is the exposure; Y is the outcome; S is an indicator of sample selection; U is an unmeasured variable.

association will generally be negative. If they are related to S in different directions, it will be positive.^{12,13} This is the source of many seemingly paradoxical results in samples biased by selection.¹⁴ An intuitive explanation of the consequence of conditioning on a collider on the measure of effect is described in the [Supplementary Appendix](#), available as [Supplementary data](#) at *IJE* online.

Endogenous selection bias can be generated from three general sources related to study design and analysis. When defined as sample selection, it can occur at study entry from non-participation or participants who are not equivalent to a random sample of the target. It can also occur at an intermediate level between entry and analysis from, for example, informative censoring [losses to follow-up or alternatively non-response (to a scheduled visit) which has the same consequence]. Finally, when defined as conditioning on a collider in the analysis, it is generated in the analysis stage, whether: in the development of the propensity score with conditioning on what is thought to be a pre-treatment confounder but is a collider; or, in the case of mediation analysis, by conditioning on a mediator that is also a collider; or, in the regression analysis, by conditioning on time-dependent confounders that have also become colliders.

Conditioning on a collider as sample selection or as a variable in the analysis, results in the same type of bias. Here, we focus on the bias when entry or maintenance in the studied sample is not random. As mentioned above, both entry into the sample and attrition can be seen as a missing data problem.⁸

We now give a few concrete examples of endogenous selection bias from sample selection. It can occur at entry in a cohort study as demonstrated by Pizzi *et al.*,¹³ the authors compared a time-defined regional population database of pregnancies with a self-selected (restricted population) web-based birth cohort. The selection process changes the association between the exposure variable and other covariates, and the association with outcome in comparison with the target population, and therefore changes the confounding pattern. A particular type of cohort study (called index-event study), where subjects are included based on a disease to ascertain the relationship between some exposure and a secondary outcome such as mortality, can result in paradoxical biased results.¹⁴ In this situation the exposure influences selection with the index event

disease and so do unmeasured variables that also cause outcome. An example of such results would be the protective effect of obesity on mortality among end-stage renal disease patients.¹⁵ Cross-sectional surveys can also be a source for this bias:¹⁶ for example, all eligible subjects answer a questionnaire in a field survey on HIV, but some refuse to give blood. The objective is to measure the prevalence of HIV based on blood testing and to assess the relationship with risk factors. This would be a typical incidental truncated sample situation where selected ($S=1$) and unselected ($S=0$) subjects and their characteristics are known, but only the selected have a measure of outcome. Sample selection is also known to occur in community-based case-control studies where by definition there is an outcome-dependent selection (the fraction of controls selected is different from that of cases); moreover, exposure may not be independent of (separated from) selection. The latter likely occurs more often among controls than cases because, although cases may have a direct interest in the study, control participation may seem more relevant to more highly educated subjects, resulting in a distribution of the exposure that is different from the base. Finally, selection bias can also occur from informative censoring (loss to follow-up) where censoring is dependent on exposure and a cause of the event of interest. There is bias when future risk is different between the censored and uncensored.

Target or source population, generalizability, transportability

The fundamental questions related to selection bias become: is the causal effect identified in the study sample and is it generalizable?

First, we attempt to define terms based on the recent literature. Bareinboim and Pearl² use 'study population' for the sample, and 'population of interest' for the target. Generalizability is from study population to target, and transportability is for 'extrapolation across domains (different settings, populations, environments) that differ both in distributions and inherent characteristics from the population of interest'. On the other hand, Haneuse¹⁷ uses 'study sub-sample' for the sample and 'study population' for target which is a well-defined population of interest to whom the results are intended to be generalized. Finally, Lesko *et al.*⁴ define external validity as the extent to which an internally valid effect measure in a study sample in an [...] unbiased estimator of the treatment effect in the population of interest [the target population]. External validity as generalizability is concerned with inference from a possibly biased sample of the target population back to the full target population (including the study sample). Transportability means making inference for a target population when the study sample and the target population are

partially or completely non-overlapping. Although terms from different authors are different, we can probably agree that the key ones are 'sample', and 'target population' and that generalizability means from the sample to the target from which it comes (assuming it is defined), whereas transportability is from the sample to a distinct population.

Let us consider the terms above in reference to a recent study: the study sample consisted of Manitoba (Canada) residents 40 years or older who underwent initial dual-energy X-ray absorptiometry of the spine and hip for bone mineral density, between 1999 and 2013.¹⁸ In women in particular, results showed that low body mass index and body fat percentage were associated with increased all-cause mortality. This may be an example of paradoxical results from conditioning on sample selection or possibly on an index condition¹⁵ (e.g. bone density); the criteria to identify a causal effect in the sample were possibly not met, resulting in a biased measure of effect. However, the authors assumed internally valid results, and their main conclusion underscored these findings as applicable to an undefined but more general population. The potential for selection bias is informally addressed in the discussion. Even if results were valid in the study sample, the study provided no analysis showing an effect measure generalized to a target population (external validity).

A relevant question with respect to generalizability is: can we be more specific about the target population we want to generalize to? The question is not trivial but remains ambiguous in the literature. As discussed by Lesko *et al.*,⁴ we want to generalize to 'some hypothetical population from which the study sample was randomly sampled'. This remains somewhat vague. Haneuse¹⁷ recently argued that it is essential to explicitly define the target population (which he calls the study population) for generalizability. His example is from an antidepressant study; the study population is defined as 'all adults aged 18–65 years with a diagnosis of depression at Group Health (a health plan) and who initiated a new episode of drug monotherapy between January 2006 and November 2009'. Although helpful because very specific, one could question the limited impact of generalizing to this target population. In the Manitoba study¹⁸ it could possibly be defined as Manitobans 40 years old and over in the same time period. In that case, generalizability is eminently relevant from a public health standpoint much more, could it be argued, than internal validity in this particular health-seeking sample. Therefore, the question of generalizability to the target, which has usually held second place in the minds of epidemiologists in comparison with internal validity, should probably be given much more consideration.

We would argue that more formal work is needed to define a plausible target population, given the features of the study sample. This reflection will be influenced by the feasibility of obtaining data to generalize a causal effect from

a sample to a target population, but there should be room for discussing an optimal target. For example: if one carries out a case-control study of autopsied subjects (a case is one who had died and had an autopsy) within a cohort study, should the target be those who are alive and enrolled at the time of a death/autopsy, or those enrolled at baseline or even the eligible to enter the cohort? One suggestion is to define this population based on the inclusion criteria in an emulated randomized experiment called target trial.¹⁹

Conditions to recover the causal effect with sample selection

Recently, conditions to recover causal effects with sample selection were outlined by Bareinboim and Pearl.² The authors call these conditions the ‘selection backdoor criterion’. This criterion is not fully applicable to the case-control study, which is an inherently outcome-dependent model⁸ ($Y \rightarrow S$), and we come back to this point using conditions outlined by Didelez *et al.*³

There are four conditions defining a selection backdoor criterion,² two of which are for exchangeability between exposed and unexposed and two for exchangeability between sampled and unsampled. If all four are met, the causal effect in the sample can be generalized to the target population with appropriate estimators. Assume a variable set Z that can be specifically partitioned in Z^+ containing the non-descendants of X , whereas Z^- includes the descendants of X . Assume that Z^+ is sufficient to estimate the causal effect of exposure X on outcome Y , in the sense that all paths that end with an arrow pointing to X are blocked by this variable set.

Condition 1 states that Z^+ blocks all backdoor paths from X to Y ; condition 2 states that X and Z^+ block all paths between Z^- and Y such that $(Z^- \perp\!\!\!\perp Y|X, Z^+)$ ($\perp\!\!\!\perp$ stands for independent). Two additional conditions are needed to separate the selection mechanism from Y and to make sure that the variables needed to close the paths are representative in the sample and/or are available from external data sources. Condition 3 states that X and Z must block all paths between S and Y such that $(Y \perp\!\!\!\perp S|X, Z)$; and condition 4 states that Z , X and Y are measured in the data affected by selection bias, and Z is also measured at (or known from) the population level. In the special case where Z is independent of S , then $P(Z|S=1) = P(Z)$ and we can use Z from the sample.

The conditions can therefore be summarized in less technical language by saying that controlling for confounding involves d-separating X from Y , making X and Y independent given the variable set Z , and controlling for sample selection involves separating the sampling mechanism S from Y . Generalizing then involves using a representative distribution for the variable set Z .

Take for example Figure 1d. Here the first three conditions are met with the set Z . However, Z is not independent

of selection, so the distribution of Z in the target is needed to generalize the causal effect from the sample to the target. In Figure 1e, the sets (M) , (L) or (M, L) allow the identification of the causal effect, whereas only (L) can separate Y from S . Therefore, the set L meets the first three conditions. However, L is not independent of S . To generalize to the target, data on L from the target are needed. As for Figure 1c, if one assumes this is a prospective study to estimate the relative risk or risk difference, then we observe that variable Z meets condition 1 (and the set Z^- is empty for condition 2). However, the direct path from $Y \rightarrow S$ cannot be separated and condition three is not met. A measure of the causal effect as the relative risk or the risk difference will be biased. Note though, that the occurrence of this $Y \rightarrow S$ pattern is more likely in a case-control study and if so, a conditional causal odds ratio can be recovered (discussed below). A causal effect cannot be recovered from Figure 1a and b even with the odds ratio because X has to be d-separated from S .

In a study where sampling depends only on binary Y , an estimate of the odds ratio will be unbiased (see Westreich⁸), whereas the probability of outcome as well as both the relative risk and the risk difference will be biased. This situation can describe a case-control study where there is an inherent outcome-dependent selection. From the previous statement, it will always be true that the crude marginal odds ratio (OR) in a case-control study will be unbiased, assuming no confounding and selection depending only on Y . This is a rare occurrence. For the common situations with confounders, additional conditions are needed, which are outlined by Didelez *et al.*³ To recover the conditional causal odds ratio (COR), condition 1 states that a sufficient set Z blocking all backdoor paths between X and Y is necessary. This condition is about confounding and does not involve S . Condition 2 states that $X \perp\!\!\!\perp S|(Z, Y)$. If the first condition is satisfied, it means that the observed $OR|Z = COR|Z$. If the second condition is satisfied, it means that the observed odds ratio in the selected sample (indicated with $S=1$) that is, $OR|(Z, S=1) = OR|Z$. Together the conditions mean that $OR|(Z, S=1) = OR|Z = COR|Z$. This is useful in practice, since $OR|(Z, S=1)$ can be easily estimated using only the data from a biased sample. Whereas the conditional COR can be recovered, this does not apply to the marginal OR. The marginal causal effect is obtained from estimating the following: $P(Y_x) = \sum_z P(Y|X, Z) \times P(Z)$. Whereas with outcome-dependent sampling such as $Y \rightarrow S$, the conditional COR can be estimated using the symmetry property of the OR and $P(X|Y, Z)$, the latter cannot be used in the marginal causal effect equation.

Estimators to recover causal effects with sample selection

Lesko *et al.*⁴ propose two estimators to generalize a causal effect to the target population: the g-formula, which is

equivalent to regression standardization (or its analogue stratification-based standardization with binary variables and very large samples requiring no modeling assumptions), and inverse probability of sampling weighting (IPSW). The former is an outcome regression model and the latter an exposure regression model. We limit this short development to an analytical situation with no time-varying exposures. For time-varying exposures, one needs to consider sequential standardization where outcome regression such as the g-formula is problematic, but a marginal structural model with sequential exchangeability for exposure using exposure regression with IPW would work.

Very briefly, the g-formula is:

$$\Pr(Y_x = 1) = \sum_z \Pr(Y = 1|X = x; Z) \Pr(Z)$$

where Y_x is the potential outcome under exposure level x , and $\Pr(Y = 1|X = x; Z)$ is the factual outcome (from conditional effects) that will be averaged over sufficient set Z for exchangeability. Assume binary exposure X in an observational study with sample selection and a blocking set ($S \perp\!\!\!\perp Y|X, Z$) that includes one categorical variable Z ; the average treatment effect is obtained as):²

$$\sum_z \{E[Y|X = 1, Z = z, S = 1] - E[Y|X = 0, Z = z, S = 1]\} P(Z = z)$$

where $E[Y|X = x, Z = z, S = 1]$ for $X = 1$ or 0 is obtained from the study sample. If $P(Z)$ is taken from the sample, a marginal effect is estimated in the sample; if $P(Z = z)$ is representative of the distribution in the target population, either in the sample or more commonly because it was obtained from the target population, then the causal effect in the sample can be generalized to the target. The distribution of Z from the target population provides the sampling weights. This should correct the effect for selection bias provided all assumptions for causal inference (see Lesko *et al.*⁴) are met.

A second option is an inverse probability of sampling weighted estimator (IPSW):

$$\frac{\sum_i Y_i I(X_i = 1, S_i = 1) IPW_i}{\sum_i I(X_i = 1, S_i = 1) IPW_i} - \frac{\sum_i Y_i I(X_i = 0, S_i = 1) IPW_i}{\sum_i I(X_i = 0, S_i = 1) IPW_i}$$

where $I(\cdot)$ is an indicator function such that $I(X_i = x, S_i = 1)$ is equal to 1 if $X_i = x$ and $S_i = 1$, or equal to 0 otherwise.

The IPSW estimator is obtained as a probability of selection into the study based on baseline covariates Z (meeting the criterion of separating selection from outcome) for

all possible combinations of the Z (e.g. if Z is a set of two binary variables, there would be four categories). The IPW is obtained as $1/P(S = 1|Z = z)$ with logistic regression. It is then used in the formula above to estimate the average treatment effect. Information on variables influencing selection and outcome among non-participants as well as participants is necessary to use this method. This corresponds to a missing-at-random assumption. A similar approach had been proposed by Haneuse *et al.*⁵ If there is confounding also, weights can be developed to obtain exchangeability based on exposure, and these can be multiplied by the weights creating exchangeability based on sampling.

With informative censoring, IPCW (C for censoring) can be used similarly to determine what the survival/outcome experiences of the censored participants would have been had they not been censored. IPCW creates a pseudo-population where the uncensored account for the censored with similar characteristics, thus eliminating the bias due to informative censoring. In developing the propensity score for censoring, measured common predictors of censoring and outcome are used. With this we can determine what the event experience of the censored would be had they not been censored. In a very simple example with only baseline covariates (Z), and exposure (X) and $C = 0$ for uncensored, we estimate IPCW as $1/P(C = 0|X, Z)$ or, using stabilized weights, as $P(C = 0|X)/P(C = 0|E, Z)$ which is equivalent to $IPCW \times P(C = 0|X)$. To eliminate confounding, weights for confounding are developed and used in the marginal structural model together with the selection weights.

These methods are simple but they do require access to some representative distribution in the target population for the variables involved in the selection process. Unfortunately the feasibility of that scenario may be limited because, on the one hand, one needs to know which variables influence selection and, on the other hand, measures of these are necessary.

An alternative method that does not require additional data from the target population is the Heckman sample selection model for incidental truncation (Heckit).^{20,21} Examples can be found in the epidemiological literature.^{16,22} Briefly this is a parametric sample selection model that is particularly dependent on correct specification. It requires the assumption of bivariate normality of the error terms in the selection model and the outcome model. The assumption is needed for consistent estimation of the parameters in the selection equation; it also implies a particular non-linear relationship for the effect of the variables from the selection model on the outcome through the inverse Mills' ratio. Without it, the model may give biased results. It is assumed that if unmeasured variables

determining selection are not correlated in the population, they are correlated in the sample, leading to inconsistent and biased coefficients for the exposure. In non-technical terms, the covariance between the error terms in the selection and outcome models is estimated and used to correct the coefficient of exposure in the outcome model with the selected sample. The application of the model requires information on the selected and non-selected and, to avoid multicollinearity, a so-called unique regressor associated with the selection but not directly with the outcome (a concept similar to the exclusion restriction with the instrumental variable approach). Models for binary outcomes can be used¹⁶ but the method works best with continuous outcomes.

Sensitivity analyses with user-supplied parameters

When there are no measures on the variables involved in the sample selection and no formal quantitative sensitivity analysis is performed, the use of negative outcome or exposure controls²³ is an alternative method to consider. The basic idea is to find a surrogate outcome or a surrogate exposure which, although not a possible effect of the exposure or a possible cause of the outcome, respectively, must be subject to the same bias as the studied exposure or outcome. Assuming such an analysis is feasible, because additional data on these surrogates are still needed, it can provide insight into the presence of selection bias. However, it does not provide information on its magnitude.

User-specified selection bias parameters can also be used, based on educated guesses or on other studies with good transportability features. Lash *et al.*²⁴ propose methods applicable to tabular data, showing a case-control example where the use of mobile phones and the occurrence of uveal melanoma was studied. The study had so-called partial non-participants, based on the fact that these provided some questionnaire information (but not the required measure of exposure). Assuming that the fully non-participating subjects had the same prevalence of use of mobile phones as the partial participants, the proportions of partial participants were applied to the non-participants to get expected target denominators in the 2 x 2 exposure outcome table. Cell selection proportions were then calculated and a selection bias OR estimated. The observed OR among full participants was adjusted with the selection bias OR. The assumptions involved in this type of analysis (e.g. similarities between non-participants and partial participants) need to be justified.

An extension of this tabular approach is available in STATA with the scripts *episens* and *episensi*.²⁵ *Episens* has

options for using a single deterministic user-supplied selection probability parameter or a probabilistic approach for this parameter with a range of values according to some user-specified distribution. A number of plausible distributions are available, such as triangular, trapezoidal and others. The range of values for the selection bias parameters can for example include the minimum, mode and maximum values; that the true parameter would fall in this range is probably more likely than the chance of it being the single value as in the deterministic approach. Nevertheless, selecting the distribution remains somewhat arbitrary.

These simple table methods are convivial and will even allow the analysis of combined biases (e.g. selection and misclassification); however, without additional programming, it is not possible to control measured confounders in the analysis of a specific bias. This considerably limits their usefulness. Therefore they should be considered exploratory sensitivity analyses when no study data related to selection are available.

Another method close to the IPW method used with external data has been proposed recently.²⁶ It can be applied to correct the OR in a case-control study with a biased sample (for example Figure 1a and b): each subject is weighted by the inverse of the probability of being selected according to Y, X and covariates. Assume that X, Y and Z are influential on selection but that no actual data are available on the selection probabilities and no external data are available to use as weights in a g-formula or for direct calculation of IPW as above; to apply the method, user-supplied scenarios for selection bias according to influential variables are generated. For example, plausible selection probabilities are required for exposed controls with high socioeconomic status (Z), for unexposed cases with a low socioeconomic level etc. Establishing these, the conditional probability $P(S=1|Y, X, Z)$ is calculated and subjects are weighted in the sample by $1/P(S=1|Y, X, Z)$. A weighted logistic regression analysis is applied including additional confounders not deemed involved in the selection process, to recover the causal odds ratio, assuming exchangeability on exposure.

Conclusion

Among epidemiologists, the impact of endogenous selection bias has not appeared to be a concern of equal importance to that of confounding. This probably reflects the fact that it is more difficult to understand the impact of selection bias on internal validity than that of confounding. However, despite the potential for selection bias in a study, it is not uncommon for results to be considered as though internally valid and generalized to a (usually undefined) target

population. Results contrary to biological expectation, as a consequence of conditioning on selection as a collider, have contributed to the awareness of this problem.²⁷ From a public health standpoint, the ability to generalize causal effects from both observational and randomized studies seems fundamental. The definition of the target population remains an issue for further discussion, in our opinion. To achieve generalizability, identification of causal effects in the presence of selection bias is the first step, followed by the application of methods to generalize these to the target. The latter requires data on all eligible subjects, participants and non-participants, suggesting careful study planning. Short of available data from the target population, user-supplied values for the bias parameters can be used to carry out a reasonable sensitivity analysis.

Supplementary Data

Supplementary data are available at *IJE* online.

Acknowledgment

The authors wish to thank Dr Jay Kaufman for comments on a later revised version.

Conflict of interest: None declared.

References

1. Elwert F, Winship C. Endogenous selection bias: the problem of conditioning on a collider variable. *Annu Rev Sociol* 2014;**40**: 31–53.
2. Bareinboim E, Pearl J. Causal inference and the data-fusion problem. *Proc Natl Acad Sci U S A* 2016;**113**:7345–52.
3. Didelez V, Kreiner S, Keiding N. Graphical models for inference under outcome-dependent sampling. *Stat Sci* 2010;**25**:368–87.
4. Lesko CR, Buchanan AL, Westreich D, Edwards JK, Hudgens MG, Cole SR. Generalizing study results: a potential outcomes perspective. *Epidemiology* 2017;**28**:553–61.
5. Haneuse S, Schildcrout J, Crane P, Sonnen J, Breitner J, Larson E. Adjustment for selection bias in observational studies with application to the analysis of autopsy data. *Neuroepidemiology* 2009;**32**:229–39.
6. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology* 2004;**15**:615–25.
7. Daniel RM, Kenward MG, Cousens SN, De Stavola BL. Using causal diagrams to guide analysis in missing data problems. *Stat Methods Med Res* 2012;**21**:243–56.
8. Westreich D. Berkson's bias, selection bias, and missing data. *Epidemiology* 2012;**23**:159–64.
9. Hernán MA. Invited commentary: selection bias without colliders. *Am J Epidemiol* 2017;**185**:1048–50.
10. Cole SR, Stuart EA. Generalizing evidence from randomized clinical trials to target populations. The ACTG 320 trial. *Am J Epidemiol* 2010;**172**:107–15.
11. VanderWeele TJ, Robins JM. Directed acyclic graphs, sufficient causes, and the properties of conditioning on a common effect. *Am J Epidemiol* 2007;**166**:1096–104.
12. Glymour MM. Using causal diagrams to understand common problems in social epidemiology. In: Oakes JM, Kaufman JS (eds). *Methods in Social Epidemiology*. San Francisco, CA: Jossey-Bass, 2006.
13. Pizzi C, De Stavola BL, Pearce N. Selection bias and patterns of confounding in cohort studies: the case of the NINFEA web-based birth cohort. *J Epidemiol Community Health* 2012;**66**: 976–81.
14. Choi HK, Nguyen U-S, Niu J, Danaei G, Zhang Y. Selection bias in rheumatic disease research. *Nat Rev Rheumatol* 2014;**10**: 403–12.
15. Flanders WD, Eldridge RC, McClellan W. A nearly unavoidable mechanism for collider bias with Index-Event studies. *Epidemiology* 2014;**25**:762–64.
16. Bärnighausen 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.
17. Haneuse S. Distinguishing selection bias and confounding bias in comparative effectiveness research. *Med Care* 2016;**54**: e23–29.
18. Padwal R, Leslie WD, Lix LM, Majumdar SR. Relationship among body fat percentage, body mass index, and all-cause mortality: a cohort study. *Ann Intern Med* 2016;**164**:532–41.
19. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *Am J Epidemiol* 2016;**183**:758–64.
20. Heckman J. Sample selection bias as a specification error. *Econometrica* 1979;**47**:153–61.
21. Winship C, Mare RD. Models for sample selection bias. *Annu Rev Sociol* 1992;**18**:327–50.
22. Habimana-Kabano I, Broekhuis A, Hooimeijer P. The effect of pregnancy spacing on fetal survival and neonatal mortality in Rwanda: a Heckman selection analysis. *J Biosoc Sci* 2016;**48**: 358–73.
23. Arnold BF, Ercumen A, Chung JB, Colford JM Jr. Negative controls to detect selection bias and measurement bias in epidemiologic studies. *Epidemiology* 2016;**27**: 637–41.
24. Lash TL, Fox MP, Fink AK. *Applying Quantitative Bias Analysis to Epidemiologic Data*. New York, NY: Springer, 2009.
25. Orsini N, Bellocco R, Bottai M, Wolk A, Greenland S. A tool for deterministic and probabilistic sensitivity analysis of epidemiologic studies. *Stata J* 2008;**8**:29–48.
26. Thompson CA, Arah OA. Selection bias modeling using observed data augmented with imputed record-level probabilities. *Ann Epidemiol* 2014;**24**:747–53.
27. Stovitz SD, Banack HR, Kaufman JS. Paediatric obesity appears to lower the risk of diabetes if selection bias is ignored. *J Epidemiol Community Health* 2018;**72**:302–08.