Biostatistics (2009), **10**, 1, *pp.* 17–31
doi:10.1093/biostatistics/kxn010
Advance Access publication on May 14, 2008

Adjusting for selection bias in retrospective case–control studies

SARA GENELETTI*, SYLVIA RICHARDSON, NICKY BEST

Department of Epidemiology and Public Health, Imperial College School of Medicine, London, UK
s.geneletti@imperial.ac.uk

SUMMARY

Retrospective case–control studies are more susceptible to selection bias than other epidemiologic studies as by design they require that both cases and controls are representative of the same population. However, as cases and control recruitment processes are often different, it is not always obvious that the necessary exchangeability conditions hold. Selection bias typically arises when the selection criteria are associated with the risk factor under investigation. We develop a method which produces bias-adjusted estimates for the odds ratio. Our method hinges on 2 conditions. The first is that a variable that separates the risk factor from the selection criteria can be identified. This is termed the "bias breaking" variable. The second condition is that data can be found such that a bias-corrected estimate of the distribution of the bias breaking variable can be obtained. We show by means of a set of examples that such bias breaking variables are not uncommon in epidemiologic settings. We demonstrate using simulations that the estimates of the odds ratios produced by our method are consistently closer to the true odds ratio than standard odds ratio estimates using logistic regression. Further, by applying it to a case–control study, we show that our method can help to determine whether selection bias is present and thus confirm the validity of study conclusions when no evidence of selection bias can be found.

Keywords: Conditional independence; Confounding; Directed acyclic graphs; Post-stratification; Retrospective case–control studies; Selection bias; Weighting.

1. INTRODUCTION

In epidemiology, observational studies are used to investigate the association between a set of risk factors and one or several health outcomes. To interpret their results, it is crucial to bear in mind the range of potential biases which might compromise inference (Greenland, 2005). These biases fall broadly into 3 categories, biases related to the selection of subjects into the study, biases arising from the way in which the data are apprehended (e.g. recall bias, truncation bias, measurement error) and finally bias due to confounding.

Retrospective case–control studies are by design more prone to selection bias than other epidemiologic studies. To be interpretable, they require that both cases and controls are representative of the same "target

*To whom correspondence should be addressed.

population." However, typically cases are identified either through a hospital or through a specialized registry, while controls are recruited by a complex process which involves among other things identifying the target population. The problem is compounded further by "self-selection" as participation of both cases and controls is voluntary. Thus, it is not always clear whether the study population forms a representative sample of the target population and whether the necessary exchangeability conditions between cases and controls hold.

In many cases, selection bias is not extreme enough to have an impact on inference and conclusions. However, there are circumstances under which even the best designed and run study is jeopardized by selection bias. Mezei and Kheifets (2006) show that selection bias in case–control studies can lead to overestimating the true odds ratio by up to a factor of 2. If selection bias is suspected, there are circumstances under which it is possible to attempt to adjust for it. The aim of this paper is to address these issues for retrospective case–control studies. However, the methods developed can be adapted to other types of studies investigating exposure—disease associations or even to survey-based studies.

Formally, selection bias occurs when the association between exposure and outcome within the study population is different from that in the target population. Selection problems range from identifying the representative sample to recruiting it and following it up. Further, selection bias can be introduced into a study at the design stage or during implementation.

In most epidemiologic papers analyzing case–control data, selection bias is addressed in the discussion; however, assessment generally remains qualitative. This paper details how we can detect and adjust for selection bias. The method requires first that a variable (or set of variables) that is highly associated with the selection criteria and hence with the biasing process can be identified. We term this the "bias breaking" variable. Second, potential bias breaking variables must be such that their distribution can be estimated from data that are not biased, and thus additional data are necessary. Despite these stringent requirements, we demonstrate using some examples that bias breaking variables are not uncommon.

The conditions for a variable to be bias breaking are formulated in terms of conditional independences and represented by directed acyclic graphs (DAGs). First, we express selection bias in a unique way in terms of DAGs, parallelling (Hernan *and others*, 2004). Then, we set up a formal framework in which it is easy to determine under what circumstances it is possible to adjust for selection bias using a bias breaking variable.

In Section 2, we motivate the paper by means of some examples of selection bias in case–control studies. In Section 3, we introduce basic DAG and conditional independence concepts. In Section 4, we describe the idea of the bias breaking variable before formally developing the estimators that adjust for selection bias. Section 5 briefly describes the simulation studies we conducted to evaluate the performance of our methods. In Section 6, we apply the estimators to a case–control study investigating the association between a congenital malformation (Hypospadias) and various risk factors. Section 7 relates the methods we have developed to post-stratification (PS) and inverse probability weighting (IPW). In Section 8, we make some concluding remarks and point to future work.

2. MOTIVATING EXAMPLES

EXAMPLE 2.1 Hospitalization bias, also known as Berkson's bias, has been extensively studied in the epidemiologic literature (Schwartzbaum *and others*, 2003). This type of bias arises when the exposure is a medical condition and hence also a reason for hospitalization and only hospital-based controls are used. If the rates of hospitalization for the 3 medical conditions (cases, exposure, and control selection criteria) are different, a spurious association can be estimated between the exposure and the disease (see Kleinbaum *and others*, 1982, for example).

EXAMPLE 2.2 In 1978, a controversy was sparked by Horwitz and Feinstein (1978) who claimed that case–control studies that had found an association between oestrogen use and endometrial cancer were dramatically overestimating the effects of oestrogen use. They suspected case selection bias, due to the fact that the cases were mostly women who had been diagnosed with endometrial cancer after they had gone to the doctor as a consequence of vaginal bleeding. As vaginal bleeding was a symptom of oestrogen use, women who took oestrogen could be overrepresented, thus inducing a spurious association between oestrogen use and endometrial cancer. The controversy was eventually decided in favor of the effect of oestrogen use. However, this showed that selection bias can affect case as well as control selection.

EXAMPLE 2.3 A typical problem in population-based case–control studies is that control selection is biased by the socioeconomic status (SES) of the controls. It is often found that controls with higher SES are more likely to respond than those with lower SES. Mezei and Kheifets (2006), henceforth MK, consider a situation where there is differential selection of cases and controls in different SES levels. In a meta-analysis of studies investigating the relationship between childhood leukemia and exposure to magnetic fields (EMF), MK noticed that in studies where a questionnaire and a home measurement of EMF levels were required, the participants that allowed a home measurement were usually those with higher SES, and hence those with potentially lower EMF readings since more affluent individuals are less likely to live close to sources of EMF, such as overhead power lines, than those with low SES. Case selection bias associated with levels of SES is less likely to be a problem as, typically, cases are eager to participate. Hatch *and others* (2000) investigate the possibility of bias due to selection in childhood leukemia and EMF studies, using the complete data with logistic regression methods. They find some bias due to differential selection.

From the examples described above, we see that selection bias can occur in the design stage of a study (Examples 2.1 and 2.2) or in the data-gathering stage (Example 2.3). However, in retrospective case–control studies, adjustment for selection bias can only be made during the analysis.

The problem of selection bias can be seen as a problem of exchangeability. Essentially, the case and control populations cannot be assumed to be drawn from the same (target) population. Thus, they are not exchangeable conditional on their case/control status and the underlying distribution of the exposure is not the same in the study and target populations. In the case of hospitalization bias (Example 2.1), the different rates of hospital admission of cases and controls makes them nonexchangeable with the target population. In Example 2.2, the study population has a different distribution of vaginal bleeding, and hence oestrogen use than the target population. Finally, the study and base populations in Example 2.3 have different distributions of SES and hence potentially different exposure to EMF.

3. KEY CONCEPTS

In this section, we describe selection bias in terms of conditional independences and DAGs. The DAG framework provides an intuitive context in which to express selection bias in case–control studies and determine potential sources. First, we introduce the machinery and the concepts required.

For the remainder of the paper, unless otherwise specified, the variable for the exposure is denoted by W and the disease or outcome by Y. Both are assumed to be binary. The variable representing whether a unit is selected into, or participates in, a case–control study is denoted by S and is also binary.

We use the notation $A \perp\!\!\!\perp B|C$ (Dawid, 1979) to signify "A is independent of B given C," where A, B, and C are random variables. A DAG is "a graph made up of nodes connected by directed edges (arrows) such that there are no cycles and no edges from a node to itself" (Lauritzen, 1996). DAGs can

be used to encode conditional independence structures; conversely, these can be read off a DAG using the "moralization criteria" (Lauritzen, 1996).

3.1 *Selection bias in terms of DAGs*

In order to understand how selection bias can be expressed in terms of conditional independences in a DAG, consider Examples 2.1–2.3, represented by DAGs in Figures 1(a–d).

First, consider Example 2.1. We want formally to express the ideas that (i) in the target population Y and W are not associated but that (ii) after selection into study population they are associated. These ideas can be expressed, respectively, as conditional (in)dependences (3.1) and (3.2):

$$Y \perp\!\!\!\perp W, \tag{3.1}$$

$$Y \not\perp\!\!\!\perp W | S = 1. \tag{3.2}$$

The DAG in Figure 1(a) represents conditional independences (3.1) and (3.2). It is the simplest expression of selection bias with 2 directed edges pointing from W and Y into S in a "v" shape termed a v-structure.

In Example 2.2, the exposure and the disease are associated. However, this association is distorted because the selection criteria favor women who have vaginal bleeding (B), a symptom of oestrogen use (W). Depending on whether vaginal bleeding is (i) not associated with endometrial cancer or (ii) associated with endometrial cancer (for instance it might be symptom), we have 2 ways of encoding the problem in terms of conditional independences. If (i), then

$$Y \perp\!\!\!\perp B | W, \tag{3.3}$$

$$W \perp\!\!\!\perp S | (Y, B). \tag{3.4}$$

One of the 3 possible DAGs encoding (3.3) and (3.4) is shown in Figure 1(b). If (ii) is the case, then only conditional independence statement (3.4) holds and (some) associated DAGs are given in Figures 1(c) and (d). These 2 DAGs are said to be "Markov equivalent."

Consider Example 2.3, where the exposure and the disease are again associated, but the SES B is associated with selection and is also a potential confounder. The conditional independence that describes this scenario is again (3.4) with associated DAGs in Figures 1(c) and (d). However, the role of B is different in the 2 examples. The 2 scenarios can only be distinguished from one another by introducing an additional variable (Dawid, 2002; Geneletti, 2005, 2006) such as an intervention on the exposure W.

All the DAGs in Figure 1 have a common element, namely, that there is a v-structure from W and Y to S when we "collapse" over the remaining variables. This is the key feature in selection bias formalized in Section 3.2. DAGs that are Markov equivalent to those we consider above are given in Section 4 of the supplementary material, available at *Biostatistics* online.

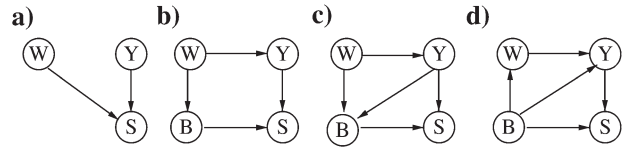

Fig. 1. (a) DAG representing selection bias without exposure and disease association, (b–d) DAGs with selection bias when exposure and disease are associated.

3.2 *Odds ratios*

The aim of inference in case–control studies is to estimate the "true" odds ratio ψ of disease given exposure

$$\psi = \frac{p(Y=1|W=1)p(Y=0|W=0)}{p(Y=1|W=0)p(Y=0|W=1)} = \frac{p(W=1|Y=1)p(W=0|Y=0)}{p(W=1|Y=0)p(W=0|Y=1)}$$

$$= \frac{\pi_1 \times (1-\pi_0)}{(1-\pi_1) \times \pi_0}, \tag{3.5}$$

where $p(W=1|Y=y) = \pi_y$. What is actually computed is ψ^o given in (3.6), which is biased if there is an association between exposure W and selection criteria S, that is, when $p(W=1|Y=y) \neq p(W=1|Y=y, S=1)$ for $y \in \{0, 1\}$

$$\psi^o = \frac{p(Y=1|W=1, S=1)p(Y=0|W=0, S=1)}{p(Y=1|W=0, S=1)p(Y=0|W=1, S=1)}. \tag{3.6}$$

4. The bias breaking model

The basic idea behind the bias breaking model is as follows: Suppose that a case–control study is suffering from selection bias because the selection criteria are associated with the exposure. However, the 2 are not associated in an obvious way, otherwise this could have been take into account when planning the study. Rather, there is a variable (or set of variables) associated with the exposure that is influencing the selection rates in a way that is either impossible to control for (such as self-selection) or unexpected. If this variable is such that it somehow "separates" the exposure from the selection criteria, then under certain circumstances detailed below, we can adjust for selection bias. This variable is termed the bias breaking variable and denoted by B.

For the sake of simplicity, we concentrate on the situations where there is an association between the exposure W and the disease Y. However, the estimators developed in Section 4.3 can be used to adjust for selection bias whether or not there is an association, as we show by simulation in Section 5. We also assume that the bias breaker B is discrete or, if it is continuous, can be appropriately stratified.

Assumption 4.1 Case and control selection are independent processes and thus can be treated separately.

This is usually plausible as cases and controls are recruited in different ways. The concept of "separation" can be formalized in terms of conditional independences and is the second assumption on which the bias breaking model is based.

Assumption 4.2

$$W \perp\!\!\!\perp S|(Y, B). \tag{4.1}$$

This is the assumption on which our method hinges: conditional on disease status, within strata of B the exposure is not associated with the selection criteria. Equation (4.1) has a simple consequence

$$p(W=1|Y=y, B=i, S=1) = p(W=1|Y=y, B=i), \tag{4.2}$$

where B is assumed to be discrete taking on values $i = 1, \ldots, n$ and $y \in \{0, 1\}$. Equation (4.2) means that we can estimate the unbiased right-hand side using an estimate of the left-hand side which is the observed

B stratum-specific proportion of exposed cases or controls in the study. Thus,

$$p(W=1|Y=y) = \sum_{i=1}^{n} p(W=1|Y=y, B=i) \times p(B=i|Y=y)$$
$$= \sum_{i=1}^{n} p(W=1|Y=y, B=i, S=1) \times p(B=i|Y=y). \quad (4.3)$$

From (4.3), we see that if we can estimate $p(B=i|Y=y)$, then it is possible to estimate the true $\pi_y = p(W=1|Y=y)$ and hence the true odds ratio ψ by using the study data. The question then is how to estimate $p(B=i|Y=y)$.

Note that based on the assumption in (4.1) only, B is a confounder for the effect of W on Y. If B is not a confounder, then (3.3) holds as well. Finally, we require the following assumption.

ASSUMPTION 4.3 The bias breaker, B, is such that additional data are available, so that we can obtain a bias-corrected estimate of its distribution, $p(B=i|Y=y)$.

This is a necessary assumption, as in order to estimate $p(B=i|Y=y)$, we need data that are not subject to selection bias. Thus, additional data must be found. These can be other data gathered within the study that contain appropriate "partial information" on B (see Section 4.2 below) or data that are external to the study itself.

Although we only consider B discrete above, the setup can be extended to consider a continuous B, where $p(W|Y, S=1, B)$ is a continuous function of B. We then need to estimate the density of B conditional on Y.

Note that once we have an adjusted estimate of π_y, we can compare this to the naive estimate $p(W=1|Y=y, S=1)$ which uses only the study data itself. If these are significantly different, there is evidence of selection bias mediated by B.

4.1 *Conditional versus marginal estimators*

When the disease under investigation is rare, as in Example 2.3, and there is only control but not case selection bias, then often the marginal distribution $p(B=i)$ is a good approximation to the conditional distribution $p(B=i|Y=0)$ required in (4.3).

Thus, another way of estimating (4.3) is

$$p(W=1|Y=y) \approx \sum_{i=1}^{n} p(W=1|Y=0, B=i, S=1) \times p(B=i). \quad (4.4)$$

In Section 4.3, we look at adjusted estimators based both on the conditional $p(B=i|Y=y)$ and marginal $p(B=i)$ distributions of B. The marginal approximation given in (4.4) is not usually appropriate if there is case selection bias.

4.2 *Additional data sources*

EXAMPLE 4.4 In Example 2.1, selection bias comes about because controls are selected among people who have been hospitalized for one or more medical conditions (C), generally chosen to be unrelated to the disease under investigation (Y). The bias breaking variable in this situation is therefore the hospitalization H given the condition C. Thus, we must estimate $p(H, C|Y)$ to adjust for selection bias. When the disease is rare, we can approximate $p(H, C|Y=0)$ with $p(H, C)$, the population rather than control distribution.

The additional data needed to do this can be found in large government databases. In the United Kingdom, there are 2 such sources: the Hospital Episode Statistics database and the Health Survey for England.

EXAMPLE 4.5 In Example 2.2, the problem is one of case selection and the bias breaking variable is vaginal bleeding V. The probability needed to adjust for bias is $p(V|Y = 1)$, which can be estimated by the proportion of women (in the population) with endometrial cancer who experience vaginal bleeding. As endometrial cancer is such that almost all women with the condition are eventually identified, additional data in the form of registry and medical records can be used to get a handle on $p(V|Y = 1)$. These data are external to the study itself.

EXAMPLE 4.6 Consider the studies on the association of childhood leukemia and EMF in Example 2.3. In most studies (see Mezei and Kheifets, 2006), analysis is conducted using only data on "full" participants, that is, those who completed detailed questionnaires and allowed magnetic field measurements (the exposure of interest) within their homes. The partial participants who only completed the questionnaire are excluded. Selection bias is suspected to enter these studies precisely because people with lower SES are less likely to allow measurements within their homes. If we assume that SES is the bias breaker B and pool the SES data from the questionnaires of the full and partial participants, we can obtain an estimate of B's distribution among controls $p(B|Y = 0)$. In this situation, the additional data have been collected as part of the study itself. In Section 6, we fully develop a similar example.

Examples 4.4 and 4.5 above are examples of "evidence synthesis" (Ades and Sutton, 2006). This term is used to describe analyses where information from different sources is combined to make better inference. When combining data to adjust for selection bias using a bias breaking variable, it is necessary to carefully assess whether synthesis is appropriate.

4.3 *Adjusted estimators*

In this section, we present our proposed selection bias-adjusted estimators. We look at both conditional estimators based on (4.3) and marginal estimators based on (4.4). The estimates of the distribution of the bias breaker can be seen as weighting the study estimates by the stratum-specific exposure probabilities.

Notation. The notation in the paper is somewhat complex because we need to index both the types of estimator and additional data available, alongside the strata and the case–control status, so we give a brief overview. Our focus is on estimating the true (conditional or marginal) distribution of B, denoted by θ. Estimates of θ are denoted by $\hat{\theta}$ with relevant superscripts which indicate the nature of the estimate—whether it is conditional or marginal and so on—and subscripts indicating the stratum of B.

We also need to distinguish between different sources of additional data. We focus on 2 types. The first is partial participant data which we term internal data. The second is data that are external to the study itself such as census data termed external data. We mention both types in Example 4.6 above.

In the internal case, we observe the columns C1–C4 of the table below in each stratum i of B as well as the respective totals over strata of B in columns C5 and C6. In the external case, we observe columns

		C1 $W = 0$	C2 $W = 1$	C3 Full study	C4 Partial	C5 Study Total	C6 Part Total	C7 External	C8 Ext Total
Y	0	K_i^0	K_i^1	K_i	K_i^p	K	K^p	K_i^*	K^*
	1	D_i^0	D_i^1	D_i	D_i^p	D	D^p	D_i^*	D^*
	Totals	N_i^0	N_i^1	N_i	N_i^p	N	N^p	N_i^*	N^*

C1–C3 as well as C5 and C7 of the table below in each stratum of B and the respective totals over strata of B in column C8.

Estimators of π_0. For didactic purposes in this section, we consider a study which has only control selection. We thus focus on estimating $\pi_0 = p(W = 1|Y = 0)$, the control exposure probability, by means of different estimates of θ. For the sake of comparison, the "naive" estimate of the probability of the exposure for controls is

$$\hat{\pi}_0^{\text{nv}} = \frac{K^1}{K} = \frac{\sum_{i=1}^n K_i^1}{\sum_{i=1}^n K_i}. \tag{4.5}$$

The general expression for the estimate of the probability of exposure in controls adjusted for selection bias, $\hat{\pi}_0$, is

$$\hat{\pi}_0 = \sum_i^n \left(\frac{K_i^1}{K_i} \times \hat{\theta}_i \right), \tag{4.6}$$

where K_i^1/K_i is the proportion of exposed controls within the $B = i$ stratum and $\hat{\theta}_i$ is the generic estimate of $p(B = i|Y = 0)$. In Table 1, we show the expressions for $\hat{\theta}_i$ for different sources of additional data. By plugging (4.7–4.10) for $\hat{\theta}_i$ into (4.6), we obtain a variety of expressions for $\hat{\pi}_0$ which cater to different additional data source and selection bias assumptions.

Adjusted estimates for $\hat{\pi}_1$ can be derived in similar ways if we suspect instead that it is the selection of the cases that are biased. Further, estimates can be derived to adjust for both case and control selection bias. Note that the bias breaker need not be the same for cases and controls. This means that the proposed estimators cater for complex situations involving multiple sources of bias. Where either cases or controls have no selection bias, then the simple (naive) estimate can be used: D^1/D or K^1/K. Finally, by replacing the parameters π_0 and π_1 by their adjusted estimates in (3.5), we get a bias-adjusted estimate of the odds ratio.

We have thus derived a series of estimators which depend on the type of additional data that are available, and on the assumptions we are willing to make about the source and nature of the selection bias. However, this list is by no means exhaustive. Other estimators can be developed to suit individual

Table 1. *The types of adjusted estimators of $p(B|Y)$ with examples*

Estimator	$\hat{\theta}_i$	Example
Conditional internal	$\hat{\theta}_i^{\text{c}} = \frac{K_i + K_i^p}{K + K^p}$ (4.7)	Example 4.6 where the additional data are data on the partial participants and we do not want to assume that $p(B\|Y) \approx p(B)$.
Conditional external	$\hat{\theta}_i^{\text{c}^*} = \frac{K_i^*}{K^*}$ (4.8)	Example 4.5. In this situation, we know the case status of the patients from the cancer registries or medical records, and these data are external to the study itself.
Marginal internal	$\hat{\theta}_i^{\text{m}} = \frac{N_i + N_i^p}{N + N^p}$ (4.9)	Example 4.6 if the control study data $K + K^p$ are sparse. In this case, we can combine the SES data on all the participants (both cases and controls) to obtain a less variable estimate.
Marginal external	$\hat{\theta}_i^{\text{m}^*} = \frac{N_i^*}{N^*}$ (4.10)	Example 4.4 where we make use of the large government databases and we do not know the case/control status of the individuals in the database.

contexts by using the machinery we have developed. Further, if the bias breaking variable is in fact a set of variables, the method can be extended in the obvious way.

Approximations for the variance estimates of the adjusted estimators $\hat{\pi}_0$ and $\hat{\pi}_1$ as well as derivations for the variance of the adjusted log odds ratios based on (4.7–4.10) are given by (1.3) and (1.19) in Section 1 of the supplementary material, available at *Biostatistics* online.

5. Simulations

The aim of the simulations detailed below is to create case–control study data sets with selection bias to study the performance of our adjusted estimators. We ran 2 types of simulation studies, both based loosely on Example A in MK. In this example, MK showed that in a population with no association between the disease and the exposure (i.e. $\text{OR}_{\text{TRUE}} = 1$) divided into 3 SES groups such that 20%, 60%, and 20% are in the high, medium, and low SES groups, respectively, varying the exposure and selection probabilities of the low SES group suffices to bias the estimate of the odds ratios up to 1.6.

The first simulation study assesses the performance of the estimators that use full and partial internal data, $\hat{\theta}^{\text{m}}$ and $\hat{\theta}^{\text{c}}$ (Example 4.6). The second simulation study assesses the performance of the estimators that use study and external data sources $\hat{\theta}^{\text{m}^*}$ and $\hat{\theta}^{\text{c}^*}$ (Examples 4.4 and 4.5).

In both simulation studies, we considered the simple case (i) where there is no association between exposure and disease, that is, $\text{OR}_{\text{TRUE}} = 1$, (ii) the situation when there is an association, but the bias breaking variable is not a confounder ($\text{OR}_{\text{TRUE}} = 2$), and finally (iii) the case when the bias breaking variable is also a confounder ($\text{OR}_{\text{TRUE}} = 2.41$). We chose an odds ratio of 2.41 so as to make results approximately comparable to the no-confounding scenario where $\text{OR}_{\text{TRUE}} = 2$. We looked at 3 biasing scenarios as well as 4 exposure probabilities for the low SES group.

For each odds ratio scenario, we compared the performance of the adjusted estimators to the estimator based on the coefficient of the exposure W in a multivariable logistic regression

$$\text{logit}(p(Y = 1|W = w, B = b)) = \beta_w w + \beta_b b, \tag{5.1}$$

denoted by $\hat{\beta}_w$. Furthermore, we simulated a data set that was not subject to selection bias and estimated $\hat{\beta}_w$ using this data set. We refer to this estimator as the benchmark estimator. We compare our estimators to estimates of $\hat{\beta}_w$ as this is the standard approach used in epidemiology when a variable is thought to be a source of confounding bias. Note that $\hat{\beta}_w$ is an estimate of the odds ratio conditional on B.

Each simulation was repeated 1000 times and the reported estimates of both the means and confidence intervals are averages over the replicates. The empirical standard deviation of simulation results ranges from 0.015 for the marginal external estimator (which used the most data) to 0.021 for the internal conditional estimator (which used the least data). For $\text{OR}_{\text{TRUE}} = 2.41$ in the highest selection bias situation, at least 81% of the 95% confidence intervals (computed using the variance formulae in supplementary material Section 1, available at *Biostatistics* online) contained the true odds ratio; in the lowest selection bias scenario, this was as high as 96%. Additional details of the simulation study are given in Section 2 of the supplementary material, available at *Biostatistics* online. The most relevant results are shown and discussed in Section 5.1 below.

5.1 *Results*

The adjusted estimates, in particular those based on θ^{m} and θ^{m^*} (the marginal estimators), were consistently closer to the true odds ratio than the standard estimates in both simulation studies. The multivariable logistic regression estimator, $\hat{\beta}_w$, performs better than the naive estimate but is outperformed by the adjusted estimates.

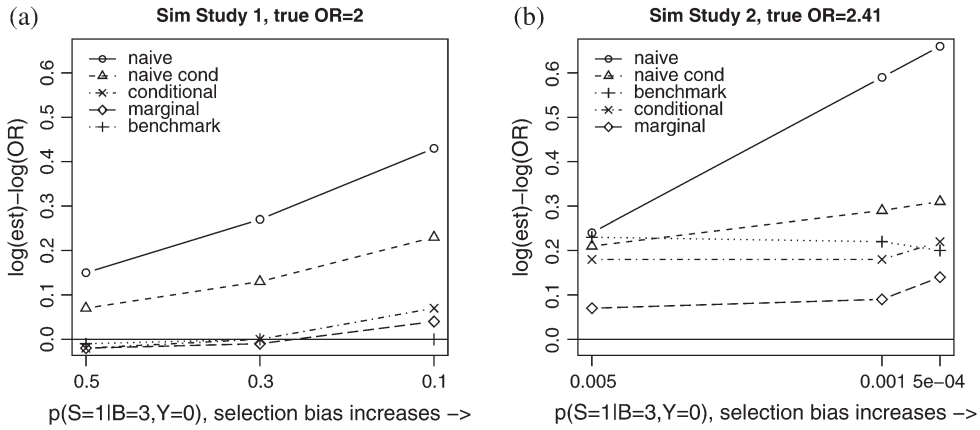

Fig. 2. Difference between the log of the estimated odds ratio and the log of the true odds ratio for simulation studies 1 and 2. The key for (a) is "naive" from a 2 × 2 table, "naive cond"$= \hat{\beta}_w$ using biased data, "conditional" = odds ratio based on θ^{c}, "marginal" = odds ratio based on θ^{m}, and "benchmark" $= \hat{\beta}_w^b$ using unbiased data. The key for (b) is conditional = odds ratio based on θ^{c^*}, marginal = odds ratio based on θ^{m^*}, while the remainder are as in (a).

Figure 2 shows the results for both simulation studies when the exposure probability is highest in the most deprived group, which leads to the most selection bias. The plot on the left-hand side of Figure 2 shows the difference between estimators and the true odds ratio on the log scale, when $\mathrm{OR}_{\mathrm{TRUE}} = 2$ in study 1, whereas the right-hand side of Figure 2 shows the differences for $\mathrm{OR}_{\mathrm{TRUE}} = 2.41$ in study 2. Selection bias increases from left to right. In both cases, the naive estimates increase, the benchmark estimate is stable, and the adjusted estimators outperform both naive estimates. These results are typical of all scenarios in both simulation studies.

Table 2 shows the odds ratio and 95% confidence interval estimates for the most extreme biasing case for all 3 odds ratios considered. The point estimates of the best adjusted estimates perform better than those of the best standard estimate. Similar results hold for the other biasing situations.

The marginal estimators in both studies outperformed the conditional estimators because they use more data—the conditional estimator is restricted by case–control status. As study 2 is intended to emulate the situation where census data are used to adjust for selection bias, it is unlikely that the case/control status of the census individuals will be known and conditional estimators would not be used.

Table 2. *Best adjusted estimators for the most extreme bias situation (*$p(W = 1|B = 3) = 0.16$*) for all odds ratios in both simulation studies. The confidence intervals were based on an approximation to the variances derived in Section* 1 *of the supplementary material, available at Biostatistics online*

$\mathrm{OR}_{\mathrm{TRUE}}$	Estimator	Simulation study 1		Simulation study 2	
		$\hat{\mathrm{OR}}$	95% CI	$\hat{\mathrm{OR}}$	95% CI
1	Best adjusted ($\hat{\theta}^{\mathrm{m}}$, $\hat{\theta}^{\mathrm{m}^*}$)	1.02	(0.45, 2.30)	1.03	(0.47, 2.27)
	Best standard ($\hat{\beta}_w$)	1.23	(0.63, 2.37)	1.22	(0.64, 2.32)
2	Best adjusted ($\hat{\theta}^{\mathrm{m}}$, $\hat{\theta}^{\mathrm{m}^*}$)	2.09	(0.95, 4.59)	2.04	(0.90, 4.65)
	Best standard ($\hat{\beta}_w$)	2.54	(1.40, 4.62)	2.52	(1.38, 4.57)
2.41	Best adjusted ($\hat{\theta}^{\mathrm{m}}$, $\hat{\theta}^{\mathrm{m}^*}$)	2.73	(1.23, 6.04)	2.76	(1.26, 6.05)
	Best standard ($\hat{\beta}_w$)	3.28	(1.85, 4.93)	3.28	(1.83, 5.92)

We derived approximate expressions for variances of the adjusted estimates (see Section 1 of the supplementary material for details, available at *Biostatistics* online). The approximation uses a specific conditional independence assumption. When there is selection bias as in the simulation studies, the independence does not hold and the variance is overestimated. Nevertheless, as a conservative guideline, we report the average size of the confidence intervals in Table 2; see Sections 6 and 8 for further discussion. Future work involves a Bayesian approach to this problem where variance estimates as developed here will not be necessary.

6. Application

The application we consider is a case–control study investigating the association between Hypospadias, a minor urogenital congenital malformation affecting baby boys which is developed during gestation, and various risk factors (Nelson, 2002; Ormond *and others*, 2007). In the study, the average income of controls was slightly higher than that of cases. This gave rise to concern about selection bias brought about by differential enrollment into the study due to SES. We thus assume that SES is the bias breaking variable.

Women in the study were administered a questionnaire that covered a range of risk factors including occupational, lifestyle, and health-related exposures as well as confounders. We only consider the risk factors: smoking, maternal age, preterm birth, all of which have been linked to Hypospadias (Porter *and others*, 2005). A detailed description of the data collection as well as variable codings can be found in Section 3 of the supplementary material, available at *Biostatistics* online.

We used 1991 ward level Carstairs score (Carstairs and Morris, 1991) standardized to cover the study region as a measure of SES. The Carstairs score is an area-level index of deprivation.

Due to the nature of the data collection process, we had access to 2 sources of data. The first was the case–control study itself (see details below). The second was the population of women of childbearing age (15–49 years) in each ward in the study area taken from the 1991 census. Using these data, we were able to estimate the distribution of SES for these women. The census data are external to the study, so we use them to calculate an additional marginal estimate for the odds ratios. We discretized the Carstairs score to 3 categories: high, medium, and low.

6.1 *Adjusted estimators*

We consider first the estimators based on the data collected during the case–control study itself. The protocol was such that the 1991 wards of residence were known for all but a small percentage of cases. Thus, even when a case did not complete a questionnaire their Carstairs score was known. The eligible controls were contacted via their general practitioner. They could reply to the organizers and decline to participate, becoming "partial" participants as their 1991 ward of residence was known but no questionnaire was completed. If they agreed and completed a questionnaire, they became full participants, for whom both the 1991 ward was known and questionnaires obtained. Finally, they could ignore the request and become nonparticipants. Due to nonparticipation, there was the possibility of additional selection bias. However, in the first part of this analysis, we assume that the pooled sample of full and partial participants is representative. The validity of this assumption is investigated when we consider using external data sources below.

Figure 3(a) shows that the partial cases have a higher Carstairs score, and are therefore more deprived, than the other subgroups. Due to their small numbers (see supplementary material Section 3.1, available at *Biostatistics* online); it was unclear whether they were a representative sample. It was thus relevant to investigate the existence of selection bias mediated by SES in both cases and controls.

Figure 3(b) shows the naive and adjusted estimates based on the internal data and their 95% confidence intervals for the 3 risk factors we considered (see Section 3.2 of the supplementary material for a

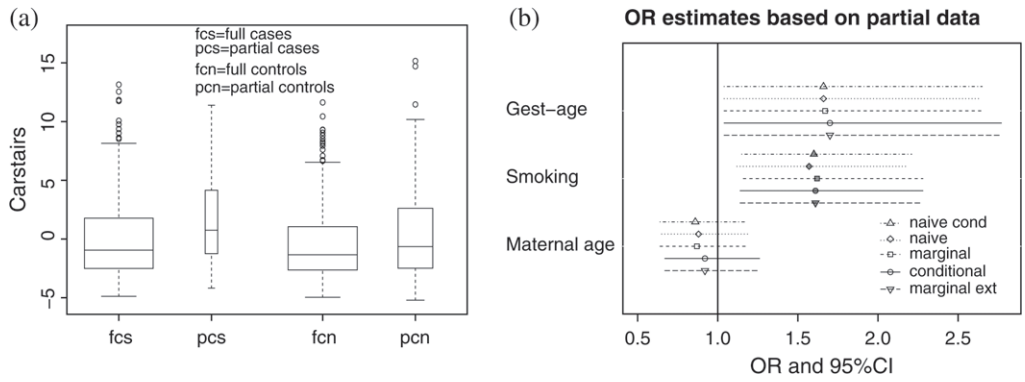

Fig. 3. (a) A boxplot of the distribution of Carstairs score for the 4 participation/status groups. The thickness of the boxes is proportional to the number of people in each group. (b) A plot of the odds ratio estimates with 95% confidence intervals for 3 risk factors. The estimate of $\hat{\beta}_w$ is labeled naive conditional, the naive estimate based on the 2 × 2 table is labeled naive, the estimates based on the internal estimators $\hat{\theta}^{\mathrm{c}}$ and $\hat{\theta}^{\mathrm{m}}$ are labeled conditional and the estimates based on $\hat{\theta^{\mathrm{m}^*}}$ are labeled marginal ext.

detailed derivation of the estimators in this context, available at *Biostatistics* online). There is practically no difference between the 4 estimates for any of the risk factors. This indicates that there is no selection bias mediated by SES, thus confirming the validity of the case–control study and its conclusions.

Note also that the variances of all the estimates are very similar. This is in contrast to the simulation studies where the variances of the adjusted estimates are noticeably larger than those of the standard estimates. This confirms indirectly that there is no selection bias mediated by SES. Indeed, when there is no selection bias, the conditional independence assumption which simplifies calculations of the variances does hold, and there is no overestimation.

In order to use the full and partial data estimators, we assumed that the eligible controls that participated were a representative sample of the base population of women living in the area covered by the case–control study, that is, we assumed that there was no further selection bias due to nonparticipation. This meant that the complete respondent data provided us with a good estimate of the distribution of SES.

We can test the validity of this assumption by using the regional distribution of the Carstairs score to estimate a bias-corrected odds ratio using the external marginal estimator $\hat{\theta}^{\mathrm{m}^*}$ and comparing it to the internal conditional and marginal estimators which use only participant data. The estimates based on $\hat{\theta^{\mathrm{m}^*}}$ are also in the plot in Figure 3(b). The marginal external estimates are very similar to the internal and naive estimates confirming the lack of selection bias mediated by SES.

A large divergence between the estimates would indicate that the study population is nonexchangeable in terms of SES with the population of women of childbearing age that we are using to adjust for selection bias. In the current context, this does not seem to be the case, and we must be careful not to overinterpret small differences in the estimates.

7. RELATED METHODS

PS and IPW are common weighting procedures. The former is used principally in survey literature and is rare in epidemiology (Samuelsen *and others*, 2006). The latter, or variants of it, are used in econometrics (Wooldridge, 2007) and epidemiology (Rotnitzky and Robins, 2005). Both methods are aimed at adjusting for potential biases.

PS is used to adjust for item nonresponse. PS depends on additional information being available that is external to the study. Typically, in the context of surveys, the additional data comes from a census or other administrative data and is in the form of population totals. PS estimates are mathematically analogous to the adjusted estimates proposed here, the differences being the nature of the exposures and outcomes of interest. Bayesian extensions using hierarchical models to smooth, or borrow strength, have been put forward by Gelman and Carlin (2001) and Gelman (2007). Due to the similarity between PS and our estimators, we can easily implement the Bayesian extensions.

IPW is a weighting procedure put forward by Jamie Robins in the epidemiologic literature. In IPW methods, additional information is in the form of selection probabilities as it is generally used for dealing with drop out or censoring. Thus, the selection mechanism is known or known to be of a particular form. For this reason, IPW methods not usually appropriate in the current context.

Finally, the estimators we propose are similar to those put forward by Hellerstein and Imbens (1999) in the econometric literature to deal with situations when the sample population used to estimate parameters is not exchangeable with the target population which is of inferential interest. In order to get estimates of the target population parameters, weighting procedures based on auxiliary information are used.

8. Discussion

In this paper, we have developed a conceptual framework for selection bias in case–control studies. By using graphical models and conditional independence statements, we were able to explore ways in which selection bias enters case–control studies and formally state suitable assumptions for estimation of odds ratios. In particular, we demonstrated how to construct a model which incorporates additional bias breaking variables to adjust for selection bias and explained how these data can be combined with study data to improve inference.

We considered a handful of plausible adjusted estimators; however, using the same principles, other estimators can be developed. When external data are sparse (e.g. when it is collected specifically to adjust for selection bias) and only control selection bias is suspected, then study case data can be combined with external control data on the bias breaker to estimate its distribution.

Using a simulation study, we showed that the estimators we have developed can be used successfully to adjust for selection bias. These estimators always outperform the standard estimators. Overall, the marginal estimators perform best because they use more data than the conditional estimators. We thus recommend using marginal estimators when possible.

We also showed, using a real data set, that the adjusted estimates can be used to check whether a potential bias breaking variable is indeed related to selection bias by comparing the adjusted to naive estimates. We note that adjusting for potential selection bias when it is not present does not introduce bias; in the application, the naive and adjusted estimators are virtually identical. This is reassuring and means that various potential bias breaking variables can be explored without compromising inference. Thus, the method can be used to validate the findings of retrospective case–control studies.

When no additional data are available, then the estimate of the distribution of the bias breaking variable $\hat{\theta}$ could be replaced by a plausible value or range of values. In this guise, the adjusted estimators can be used to perform sensitivity analysis and can also be seen as prior distributions in a Bayesian context.

The main problem with the adjusted estimators as they stand is that they have a variance which is larger than that of the standard estimates. The explanation for this inflation of the variance is that, in order to simplify the analytic derivation of the variances, we have made a conditional independence assumption which is unlikely to hold when there is selection bias.

The next step is to develop Bayesian hierarchical models in the spirit of PS (Gelman, 2007). This will have various advantages over the current approach. It will create a natural framework for sensitivity

analysis. It will provide realistic variance estimates without resorting to analytic approximations. Finally, it will simplify the inclusion of additional covariates.

ACKNOWLEDGMENTS

The authors acknowledge, Paul Elliott, Mark Nieuwenhuijsen, Paul Nelson, Mireille Toledano, Nina Iszatt, and Daniela Fecht for help organizing the data, Isabelle Stucker and Sylvaine Cordier of INSERM for discussion. *Conflict of Interest:* None declared.

FUNDING

Economic and Social Research Council (RES-576-25-5003 to S.G., S.R., N.B.); UK Department of Health (12167262) for Hypospadias Study.

REFERENCES

ADES, A. AND SUTTON, A. (2006). Multiparameter evidence synthesis in epidemiologic and medical decision-making: current approaches. *Journal of the Royal Statistical Society, Series A* **196**, 5–35.

CARSTAIRS, V. AND MORRIS, R. (1991). *Deprivation and Health in Scotland*. Aberdeen: Aberdeen University Press.

DAWID, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **41**, 1–31.

DAWID, A. P. (2002). Influence diagrams for causal modelling and inference. *International Statistical Review* **2**, 161–189.

GELMAN, A. (2007). Struggles with survey weighting and regression modelling. *Statistical Science* **22**, 153–164.

GELMAN, A. AND CARLIN, B. (2001). Poststratification and weighting adjustments. In: Groves, R., Dillman, D., Eltinge, J. and Little, R. (editors), *Survey Nonresponse*. New York: Wiley, pp. 289–302.

GENELETTI, S. (2005). Aspects of causal inference in a non-counterfactual framework, [PhD. Thesis]. London: University of London.

GENELETTI, S. (2006). Identifying direct and indirect effects in a non-counterfactual framework. *Journal of the Royal Statistical Society Series B* **69**, 1–17.

GREENLAND, S. (2005). Multiple-bias modelling for analysis of observational data. *Journal of the Royal Statistical Society, Series A (Statistics in Society)* **2**, 1–25.

HATCH, E., KLEINERMAN, R., LINET, M., TARONE, R., KAUNE, W., ANSSI, A., DASUL, B., ROBINSON, L. AND WACHOLDER, S. (2000). Do confounding or selection factors of residential wire codings and magnetic fields distort findings of electromagnetic field studies? *Epidemiology* **11**, 189–198.

HELLERSTEIN, J. AND IMBENS, G. (1999). Imposing moment restrictions from auxiliary data by weighting. *The Review of Economics and Statistics* **81**, 1–14.

HERNAN, M., HERNANDEZ-DIAZ, S. AND ROBINS, J. (2004). A structural approach to selection bias. *Epidemiology* **15**, 615–625.

HORWITZ, R. AND FEINSTEIN, A. (1978). Alternative analytic methods for case-control studies of estrogens and endometrial cancer. *New England Journal of Medicine* **299**, 1089–1094.

KLEINBAUM, D., KUPPER, L. AND MORGENSTERN, H. (1982). *Epidemiologic Research*. Belmont, CA: Lifetime Learning Publications.

LAURITZEN, S. (1996). *Graphical Models*. Oxford: Clarendon Press.

MEZEI, G. AND KHEIFETS, L. (2006). Selection bias and its implications for case-control studies: a case study of magnetic field exposure and childhood leukaemia. *International Journal of Epidemiology* **35**, 397–406.

NELSON, P. (2002). Geographical epidemiology of hypospadias, [PhD. Thesis]. London: University of London.

ORMOND, G., NIEUWENHUIJSEN, M., NELSON, P., IZATT, N., GENELETTI, S., TOLEDANO, M. AND ELLIOTT, P. (2007). Folate supplementation, endocrine disruptors and hypospadias: case-control study. Under review in BMJ.

PORTER, M., FAIZAN, M., GRADY, R. AND MUELLER, B. (2005). Hypospadias in Washington state: maternal risk factors and prevalence trends. *Pediatrics* **115**, 495–499.

ROTNITZKY, A. AND ROBINS, J. (2005). Inverse probability weighted estimation in survival analysis. In: Armitage, P. and Colton, T. (editors). *Encyclopedia of Biostatistics*, 2nd edition. New York: Wiley, **4**, 2619–2625.

SAMUELSEN, S., AANESTAD, H. AND SKRONDAL, A. (2006). Stratified case-cohort analysis of general cohort sampling design. *Technical Report*. University of Oslo.

SCHWARTZBAUM, J., ALBHOM, A. AND FEYCHTING, M. (2003). Berkson's bias reviewed. *European Journal of Epidemiology* **18**, 1109–1112.

WOOLDRIDGE, J. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics* **141**, 1281–1301.

[*Received June 29, 2007; first revision December 14, 2007; second revision January 22, 2008; third revision February 20, 2008; accepted for publication March 21, 2008*]