Uncovering selection bias in case-control studies using Bayesian poststratifictation

Sara Geneletti[+], Nicky Best[*] Mireille Toledano[*],Paul Elliott[*], Sylvia Richardson[*]
[+] London School of Economics, [*] Imperial College London

Type of article: Methodology
Corresponding author: Sara Geneletti,
Dr Sara Geneletti
Lecturer in Statistics
London School of Economics and Political Science
Houghton Street
London, WC2A 2AE, UK
Tel: +44 (0)20 7955 7664
Fax: +44 (0)20 7955 7416
email:s.geneletti@lse.ac.uk

Running title:Uncovering selection bias
Financial support:S. Geneletti, S. Richardson and N.Best would like to acknowledge support from ESRC: RES-576-25-0015, M. Toledano and P.Elliott would like to acknowledge support from UK Department of Health (12167262) for the hypospadias study. **Acknowledgements**: The authors would like to thank Dr. Martha Linet who kindly provided data files and coding manuals for the National Cancer Institute - Children's Cancer Group case-control study of residential exposure to magnetic fields and acute lymphoblastic leukemia in children.

Abstract
Selection bias can affect odds ratio estimation in particular in case-control studies. Approaches to discovering and adjusting for selection bias have been proposed in the literature using graphical and heuristic tools as well as more complex statistical methods. The approach we propose is based on a survey weighting method termed Bayesian post-stratification and follows from the conditional independences that characterise selection bias. We use our approach to perform a selection bias sensitivity analysis of odds ratios by using ancillary data sources that describe the target case-control population to re-weight the parameter estimates obtained from the study. The method is tested on two case-control studies, the first investigating the association between exposure to electro-magnetic fields and acute lymphoblastic leukaemia and the second investigating the association between occupational exposure to hairspray and a minor congenital malformation called hypospadias. In both case-control studies, the odds ratios were only moderately sensitive to selection bias.

KEYWORDS: Selection bias, post-stratification, Bayesian, conditional independence, case-control studies

1 Introduction

Selection bias (SB) can present a serious problem for valid odds ratio (OR) estimation in case-control studies as demonstrated in Mezei and Kheifhets(2006) [1] and Geneletti et al.(2009). [2] These are especially sensitive to SB as the sampling mechanism depends on the case/control status of the participants and participation probabilities are generally hard to estimate. SB comes about when the exposure under investigation is associated with the selection mechanism. As we typically have limited information on the distribution of the exposure other than from the study itself, we are unable to estimate the dependence between the exposure and the selection and can therefore not adjust for SB. In order to overcome this problem, we propose to introduce a set of variables B such that first, B separates the exposure from the selection, and second, such that the distribution of B can be estimated from sources of data external to the study. In so doing we shift the SB from the exposure, whose distribution cannot be estimated without bias from the study, to B, whose distribution is estimated from data external to the study and thus potentially unbiased. By using different sources of data to estimate the distribution of B we can investigate the sensitivity of the OR to different selection processes. This aspect is the novelty and strength of this approach as it encourages us to think carefully about the populations of interest and the sources of bias.

In Section 2 we describe the two case-control studies we use to assess our method. In Section 3 we derive our estimator and its Bayesian extension and discuss the sampling assumptions underlying our method. Section 4 covers the sensitivity analysis applied to the case-control studies. We finish with a discussion in Section 5.

2 Case-control studies

2.1 EMF and Childhood ALL case-control study

Extremely low-frequency electromagnetic fields (EMF) have been designated as possibly carcinogenic by the International Agency for Research on Cancer based on epidemiologic studies in children. [3]. We consider here a study investigating the link between EMF exposure from power cables and childhood acute lymphoblastic leukaemia (ALL) [4] which found little evidence of an association (OR for ALL of 1.24, 95% confidence interval (0.86,1.79) at exposures of 0.2μT or greater as compared with less than 0.065μT). In a later analysis, Hatch et al (2000), [5] suggested that there might be SB due

Income, race and urban status of EMF-ALL study

	Full(%)	Partial(%)		Full(%)	Partial(%)
income			*race*		
$\leq \$20k$	134(**12**)	119(**27**)	white	1031(**94**)	372(**87**)
$\$20k - \$39k$	381(**35**)	156(**37**)	black	22(**2**)	34(**8**)
$> \$39k$	577(**53**)	152(**36**)	other	39(**4**)	21(**5**)
urban			*disease*		
city	733(**67**)	117(**27**)	case	576(**53**)	189(**44**)
rural	359(**33**)	310(**73**)	control	516(**47**)	238(**56**)

Table 1: Table of numbers(percentages) of individuals by partial and full participant status in the EMF-ALL dataset. There are a total of 1092 (72%) full and 424 (28%) partial participants

to differential participation rates in different socio-economic strata.

Briefly, cases were contacted by the Children's Cancer Group and controls were selected by random digit dialling and matched to cases according to the first eight digits of their phone numbers, age and race. Demographic details were collected over the telephone. EMF measurements inside the residence of those who had completed the telephone interview were attempted by technicians blinded to the case/control status. Participants for whom indoor measurements were made, are termed *full* participants, whilst those for whom indoor measurements were not made, either because of refusal or because the family had moved etc. are termed *partial* participants.

In our analysis, we concentrate on three socio-economic indicators, race, annual household income (income) and whether the family lived in a city or otherwise (urban). These seemed the most important socio-economic indicators, particularly urban, as individuals living in a city will be exposed to more EMF via power cables than those living in the countryside.

Table 1 shows the numbers (%) of individuals in the socio-economic indicator groups by full and partial participant status. Full and partial participants are significantly different: half of full participants but only one third of the partial participants are in the highest income bracket, 8% of the partial sample are black versus only 2% in the full sample and urban dwellers were twice as likely to be full participants than rural dwellers. Finally, whilst the cases and controls are spread evenly amongst the full participants, there is a slightly larger proportion of controls amongst the partial participants.

If – as is plausible – EMF exposure is associated with the socio-economic status (SES) of the participants then this can result in SB.

2.2 Hairspray exposure and Hypospadias case-control study

The second application we consider is a case-control study investigating the association between hypospadias, a minor congenital malformation and occupational exposure to hairspray[6] which has been linked to hypospadias. Ormond et al (2007) [6] estimated that maternal exposure to hairspray was associated with increased risk of giving birth to a baby boy with the malformation (OR = 2.4, 95% confidence interval (1.40,4.17), adjusting for income and smoking).

The average household income of controls was slightly higher than that of cases, and the cases included a higher proportion of younger women than the controls. This gave rise to concerns about SB brought about by differential enrolment into the study due to SES and maternal age (MA) as both may be associated with occupational exposure to hairspray.

Women were initially contacted by mail and if they responded were asked to complete a telephone questionnaire. As with the EMF data, the study consisted of *full* and *partial* participants. The partial participants were those who had declined to participate in the study but responded to the initial mailing, whereas the full participants were those who completed telephone interviews which included relevant confounders.

We obtained a measure of SES, the 1991 Carstairs score (an area-level deprivation score[7]), for full and partial participants with post-codes that could be linked to appropriate wards (for details see on-line supplementary materials). The Carstairs score was discretised into three categories, high, medium and low using tertiles.

Table 2 illustrates the differences between the full and partial cases and controls. While the full cases are evenly distributed amongst the three SES groups, a lower proportion of the full controls and a higher proportion of the partial cases and controls are in the lower SES group.

3 Sensitivity analysis for case-control study OR

Selection bias, item non-response bias, drop-out bias, are the same type of bias arising in different contexts. This bias arises when the sample in the study is unrepresentative of the target population, i.e. has been sampled differentially with respect to some key variables. The most common approach

SES and Maternal Age in H-H study

	Cases		Controls			Cases	Controls
SES	Full(%)	Partial(%)	Full(%)	Partial(%)	MA	Full (%)	Full (%)
high	139(**33**)	20(**17**)	161(**37**)	80(**32**)	<25 yrs	51 (**13**)	37 (**9**)
medium	138(**33**)	40(**36**)	152(**35**)	72(**30**)	25-35 yrs	244 (**64**)	258 (**64**)
low	145(**34**)	52(**47**)	122(**28**)	95(**38**)	>35 yrs	83 (**22**)	111 (**27**)

Table 2: Table of numbers(percentages) of individuals by full and partial participant groups for SES in the H-H dataset. Also number(percentages) of women in 3 age categories (MA). The full participants included those for whom we had data on the variables of interest, smoking, maternal age, SES and occupational exposure to hairspray. There are a total of 857 (70%) full and 360 (30%) partial participants. These are fewer than the 1487 considered in[6] as it was necessary to have a correct postcode for the participants in order to link it with the bias breaking variable B discussed in Section 3.1.1.

to adjusting for this type of bias – which we term selection bias (SB) – is to use a weighting procedure.

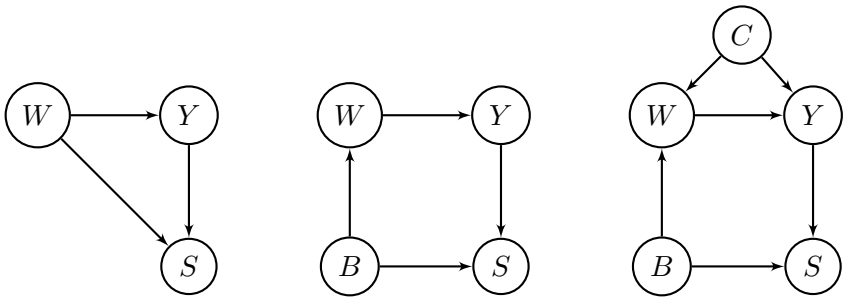
In case-control studies, cases and controls must be *exchangeable* with respect to all variables involved in the association under investigation in order to obtain valid ORs. This means that they must be sampled from the same population, e.g., if controls are predominantly office workers and cases are predominantly factory workers and we are investigating the association between exposure to a factory chemical and the incidence of a particular cancer, then we are likely to get invalid results as cases are not exchangeable with the controls with respect to employment, income etc.

The dependence between the sampling mechanism and the case/control status makes case-control studies particularly vulnerable to SB: Selection probabilities tend to differ systematically between cases and controls and are hard to estimate. This makes it difficult to assess whether the two groups are exchangeable.

A number of weighting schemes have been proposed to adjust for SB in epidemiology, mostly based on Horvitz-Thompson (HT) estimators, the most common being inverse probability weighting.[8]

3.1 Post-stratification

We focus instead on a survey weighting technique termed post-stratification (PS). PS involves weighting the parameters of interest such as outcome



(a) Selection bias (b) Bias breaking with B (c) Inclusion of confounder C

Figure 1: The idea behind the bias breaking variable approach to adjusting for selection bias.

means and probabilities by the proportion in which the associated explanatory variables occur in the population,[9] as such it is an extension of direct standardisation. PS based estimators can be derived from a simple conditional independence argument based on understanding of the SB process (see Section 3.1.1). This argument also leads to a simple way to investigate the sensitivity of the OR to different PS weights and the exchangeability assumptions they embody.

3.1.1 Conditional independence and post-stratification

SB comes about when the exposure under investigation is associated with the process of selection into a study. This means that estimates of the OR based only on the individuals for whom data are available for the exposure and the case/control status, those we termed full participants in our examples, will be biased. Let W represent the exposure of interest, Y the outcome and S whether an individual is selected into a study. In terms of conditional independence[10] we say that if $W \not\perp\!\!\!\perp S|Y$, i.e. W is not independent of S conditional on Y, then there is SB. If we use directed acyclic graphs (DAGs) to express selection bias, we see that it is a form of *collider bias* as in Figure 1a where both W and Y are directly associated with S.[2,11]

The idea underlying our sensitivity analysis is to identify a set of variables B implicated in the selection process such that conditional on B the exposure and the selection criteria are no longer *directly* associated (Figure 1b), i.e. such that:

$$W \quad \perp\!\!\!\perp \quad S|(Y, B). \tag{1}$$

This conditional independence also holds if, in addition to B there are confounders C, not implicated in the selection process. In fact, it extends to

$$(W, C) \quad \perp\!\!\!\perp \quad S|(Y, B). \tag{2}$$

with the corresponding DAG given in Figure 1c. Note that these conditions still hold if B and C are associated. We can use conditional independence (2) to decompose the probability of exposure W given case/control status, the components of the OR of interest, as follows:

$$\begin{aligned} p(W = 1|Y = y) \quad = \quad & \sum_B \sum_C \underbrace{p(W = 1|Y = y, B, C)}_{(a)} \\ & \times \overbrace{p(C|Y = y, B)}^{(b)} \times \underbrace{p(B|Y = y)}_{(c)} \end{aligned} \tag{3}$$

Equation (3) is an example of a PS equation: (a) is the parameter estimate (the B stratum specific probability of exposure given Y), (b) is the weight associated with a confounding adjustment and $(c) = P(B|Y = y) = p_{by}$ is the PS weight.

We can estimate (a) and (b) directly using data from the the full participants (i.e. those for whom $S = 1$) as conditional independence (2) allows us to replace these expressions with $p(W = 1|Y = y, B, C, S = 1)$ and $p(C|Y = y, B, S = 1)$ respectively. Crucially, we cannot similarly use $p(B|Y = y, S = 1)$ to estimate p_{by} because B its is not independent of the selection mechanism. Using only the full participant data will result in a biased estimate of p_{by}. Instead, we use other data sources to estimate p_{by} and use these estimates as tools to investigate the OR's sensitivity to SB. Our method hinges on being able to find sources of data which provide plausible estimates of p_{by} in the context under investigation. Identifying sources for plausible estimates of B is typically possible in epidemiology .[2] We refer to B as the bias breaking variable (BB).

Equation (3) highlights the difference between confounders and SB variables. While we deal with confounding by adjusting for it using the full participant data (the equivalent of adding it as a covariate in a regression), the distribution of the BB variables B needs to be estimated from additional data.

First we look at estimating (a) and (b) jointly using Bayesian techniques and then we look at how to estimate p_{by} and conduct the sensitivity analysis.

3.2 Bayesian post-stratified analysis

Bayesian post-stratification (BPS) has been used in a number of contexts [12,9] in the survey literature. Geneletti et al. (2008) [2] applied non-Bayesian PS to the H-H case-control study. We extend both approaches here.

Variances for the adjusted estimates are hard to calculate in closed form, [2] however, by using Bayesian Markov Chain Monte Carlo (MCMC) methods, the variance of the estimators is simply the empirical variance of their posterior sample. Further, MCMC gives us the distribution for the OR enabling us to investigate any aspect of the distribution.

Typically, in a frequentist analysis ORs are estimated using the maximum likelihood estimate (MLE) of the *prospective* logistic regression coefficient of the exposure, β_w in Equation (4).

$$\text{logit}\{p(Y=1|W,B)\} = \beta_0 + \beta_w W + \beta_b B + +\beta_c C \tag{4}$$

A well known result states that the frequentist MLE of β_w can be estimated from the *retrospective* logistic regression where W and Y are swapped in Equation (4). However, the aim of inference here are estimates of $p(W|Y,B,C)$ – (a) from Equation (3) – and these cannot be obtained from Equation (4).

We follow Seaman and Richardson (2001) [13] and estimate the vector of probabilities $p(W,B,C|Y=1) = \boldsymbol{\gamma}$ and $p(W,B,C|Y=1) = \boldsymbol{\phi}$ for all combinations of W,C and B, using multinomial likelihoods for $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$ with Dirichlet priors (see on-line supplementary material). From $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$, we obtain (a) and (b) from Equation (3) by probability manipulations.

The Bayesian approach has two main components: The model for the probabilities (a) and (b), and the model for p_{by}. We use independent Dirichlet priors for $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$, as more complex, e.g. hierarchical models, made little difference to the final outcomes. In both studies we used fixed weights based on the raw frequencies to estimate p_{by} as data were plentiful and a prior model was not necessary. In other scenarios, data might be sparse or particular constraints need to be placed on the weights and thus using a model to estimate the weights could be appropriate.

3.3 Sensitivity to choice of PS weights

To control for SB it is necessary to make cases and controls exchangeable with respect to variables implicated in the selection mechanism. We cannot do this directly as we typically cannot estimate selection probabilities. Thus, we assess the sensitivity of the OR by "wiggling" the distribution of $p(B|Y=y) = p_{by}$, by estimating p_{by} from different sources of data that we consider

representative of the cases and controls as defined in the study protocol. What data sources result in plausible estimates of p_{by} will depend on the source of bias and the extent to which the bias affects the cases and controls in the study – i.e. their exchangeability.

Generally, it will be appropriate to weight the cases and controls differently. Intuitively, if we have SB, this is precisely because the case and control populations are differently affected by this bias and therefore need to be adjusted for differently.

If we assume that the source of SB, if there is any, is the same for cases and controls, (as we do in our applications) then we must ask ourselves to what extent the two groups are affected. If e.g., SES is thought to be a good candidate for B, then there will often be reasons to suspect that the control sample has a larger proportion of higher SES individuals than the case sample as typically, cases will be more motivated to participate in the study irrespective of their SES. Different case-control studies will however have different BB variables and context specific reasoning on the extent of the bias will be necessary.

The sensitivity analysis we propose involves first identifying a suitable set of variables B satisfying conditional independence (2) and such that additional data (other than the full participants) are available to estimate its distribution. Second, for each additional dataset estimating at least one PS weight. Third, for each PS weight calculating a corresponding adjusted OR. The final step involves comparing adjusted ORs to assess sensitivity to choice of weight.

3.4 Sources of data to estimate p_{by}

We consider first what datasets provide sensible estimates of p_{by} and second which combinations of these estimates, for cases and for controls respectively, correspond to plausible selection scenarios and reflect our understanding of the strength of the SB in the two groups.

If the study is population based, it will often be sensible to use census or similar routinely gathered datasets to estimate PS weights as these will be a good proxy for the distribution of the BB variables in the controls. Another common situation is a study having *partial* participants, i.e. individuals for whom case/control status and covariate information are available but for whom exposure information is lacking. In such situations, it is sometimes meaningful to combine partial with full participant data to estimate *combined* participant weights. Combining is advisable when a study is well designed and conducted as the set of eligibility criteria for cases and controls

in the protocol will be such that those contacted will be representative of the target cases and controls and thus exchangeable. These will provide us with valid BB information. Combining full and partial participants is not meaningful unless the number of partial participants is a considerable percentage of the full participants (over 20%) as smaller numbers are unlikely to change results.

3.4.1 EMF-ALL

Consider the EMF-ALL example. We assume based on arguments in Section 2.1 that B, the set of socio-economic indicators given by {race, income, urban} is the BB variable for both cases and controls. One source of data to estimate p_{yb} is the Current Population Survey (CPS), a monthly survey of about 50,000 households conducted by the Bureau of the Census for the Bureau of Labor Statistics in the US. This has exhaustive information on the distribution of B in the nine States of the study.

As this study was well-designed, there is a second source of data to estimate the distribution of B: the combined full and partial participant data gathered during the course of the study for which B is known.

The diagram in Figure 2 shows the two sources of data, the combined full and partial participant data and CPS (external) and the three types of estimates of p_{by} that can be obtained from these data. In the left-hand section, we combine full and partial participants but condition on case/control status to obtain estimates of p_{by}. We term these the *combined full and partial participant weights conditional on case/control status* (CPCs and CPCn respectively). In the middle section of Figure 2 we also combine full and partial participants but this time marginalise over the case/control status to obtain an estimate of $p(B) = p_b$ rather than $p(B|Y)$. We term this the *combined full and partial participant weight marginalised over case/control status* (CPM). Finally, we consider only the external CPS data and obtain another estimate of p_b, the *external marginal* (EM) weight.

We now need to decide which weights should be combined with cases and which with controls by considering the extent of the bias in the two groups. Table 3 describes a number of plausible combinations of weights for cases and controls for the EMF-ALL study. We consider two (CPCs,CPCn) and (CPM,EM) in more detail here. The (CPCs,CPCn) estimator relies entirely on combined participant data. The assumption underlying this estimator is that the individuals *contacted* during the study following the protocol were in fact a representative sample of the target case/control population and were by design exchangeable. If this estimator is very different from the

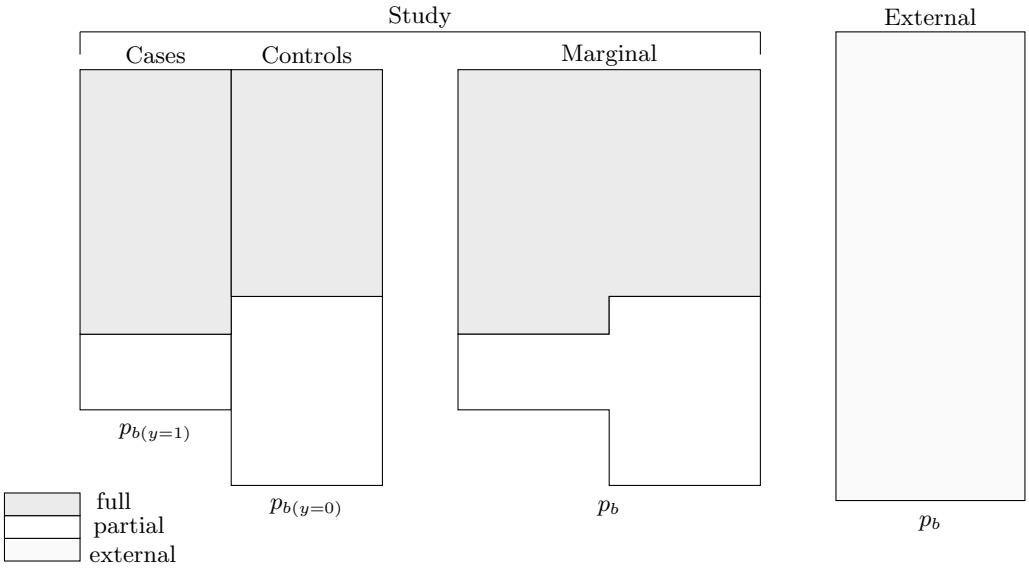


Figure 2: Diagram showing the data used to estimate p_{by}, (c) in Equation (3). The expression below each box denotes the quantity that can be estimated using the data represented by that box. We omit C for simplicity. Below each section is the quantity we can estimate using that data.

frequentist logistic regression estimate that we take as the baseline, then SB is present. The (CPM,EM) estimator is based on combined participant data for the cases and external CPS data for the controls. This estimator is plausible for the controls as they were intended to be representative of the general population, differing from the cases only in their disease status. It is a plausible estimator for the cases in this study because the raw proportions of B amongst the full cases are similar to those for the combined participants and the larger number of individuals (1519 vs 1092 full) in the combined group decreases the variance.

3.4.2 Hairspray exposures-Hypospadias

In this case-control study, we have three sources of data to draw from. The first is the 1991 Carstairs deprivation score (an area level measure of socio-economic deprivation) for the full and partial participants obtained from their electoral geographic ward of residence. Using this score we can estimate CPCs and CPCn as well as CPM weights as in the EMF-ALL study. As a quarter of the cases and a third of the controls were partial

Post-stratification weights

cas/con	Assumptions and Interpretation
CPCs,CPCn	The cases and controls contacted are representative of the target case-control sample w.r.t B.
CPCs,CPM	The distribution of B in the controls is estimated well by the combined participant data.
CPCs,EM	The distribution of B in the controls is estimated well by the CPS data.
CPM,EM	The distribution of B in the cases is estimated well by the combined participant data.
Key	
CPCs,CPCn	Estimates p_{by} using combined full and partial participant data conditioning on case/control status as shown on LHS of Figure 2.
CPM	Estimates p_b using combined full and partial participant data marginalising over case/control status as shown in middle of Figure 2.
EM	Estimates p_b using CPS data external to the study as shown on the RHS of Figure 2.

Table 3: Table describing the assumptions underlying the plausible weighting schemes for the EMF-ALL study.

participants and these complied with the study protocol, it was appropriate to use the partial information (see Table 2).

The second source of data is the the study area 1991 census. Using these data we can estimate the distribution of 1991 Carstairs score for women of childbearing age (15-55). The census data are external to the study and thus contains no case/control information, so we use them to estimate an EM weight. This is analogous to the EMF-ALL EM weight.

The third source of data is the Millennium Cohort Study (MCS) which can be used to estimate the distribution of maternal age (MA). MA is also a confounder for the association between hypospadias and hairspray and was measured in the case-control study. The MCS data provides us with an additional EM weight.

Similar arguments as those used for the EMF-ALL study can be put forward when considering what combinations of weights to use for the cases and controls in the H-H study. However in this study, we consider two rather than one set of BB variables B, $B = SES$ (as estimated by the Carstairs score) and $B = \{SES, MA\}$ where MA represents maternal age.

Adjusted odds ratios for EMF-ALL study

cas/con	Med	CI
CPCs,CPCn	1.13	(0.73,1.74)
CPM,EM	1.04	(0.69,1.62)
Frq lg	1.14	(0.77,1.68)

Table 4: Medians and credible/confidence intervals for the odds ratios of ALL for exposure to EMF ($\geq 0.2\mu$T) using (CPCs, CPCn), (CPM, EM) and logistic regression estimates.

4 Results

4.1 EMF-ALL study

We apply the BPS method to the EMF-ALL study with $B = \{$race, income, urban$\}$ using two different sources of data for the weighting: The combined data on full and partial participants to estimate CPCs, CPCn and CPM weights and the CPS data for external weights. For both datasets we use the multinomial-Dirichlet model described in Section 3.2.

Table 4 shows the medians and credible/confidence intervals for two of the plausible weightings listed in Table 3 as well as the OR estimates of the frequentist logistic regressions of the form given by Equation (4). We chose (CPCs, CPCn) and (CPM, EM) as they represented the highest and lowest PS weighted estimates. From Table 4 we see that the median of the (CPM, EM) estimate is shifted towards one with respect to the others.

Although the changes are modest and do not affect the epidemiologic conclusions of the study, there is evidence of the OR's sensitivity to different PS weights and consequently to different sampling assumptions. In a situation where the OR is not so clearly covering one, these changes might lead to a change in the interpretation of the results.

4.2 Occupational hairspray exposure - Hypospadias study

In the H-H study we can consider a number of different data sources and variables as being potentially involved in the selection process.

First we consider the simplest case where $B = SES$ as measured using the Carstairs scores from the combined full and partial study data, resulting in CPCs, CPCn and CPM weights and the census data, resulting in an EM weight.

Keeping the same $B = SES$ we then add smoking Sm as a confounder, $C = Sm$.[14] In so doing, we assume that smoking is only a confounder and not a bias breaking variable (i.e. conditional independence (2) holds for $Sm = C$). This means that we can use the full participant data to estimate $p(Sm|SES, Y)$ and then estimate $p(SES|Y)$ as detailed above. We then use Equation (3) to combine the two probabilities, $p(Sm|SES, Y)$ is (b) and $p(SES|Y)$ is (c). This gives us additional weights.

Finally, we assume that $B = \{SES, MA\}$ where MA is MA and we use the Millennium Cohort Study (MCS) to estimate $p(SES, MA|Y)$, keeping Sm as a confounder as above. See on-line supplementary material for details on how to incorporate MA. This gives us an additional EM weight.

The variable occupational hairspray exposure takes on three values, 0,1 and 2 representing no exposure, exposure and unemployment respectively. We focus on the OR of 0 vs 1.

We consider the same multinomial-Dirichlet model which we used for the EMF data. Table 5 shows the posterior estimates for the PS adjusted OR of exposure using the (CPCs,CPCn) and (CPM,EM) weights for the three data sources used to assess the sensitivity of the OR described above. These were chosen because they had the highest and lowest medians respectively, however similar trends are true for the other PS weighted estimates.

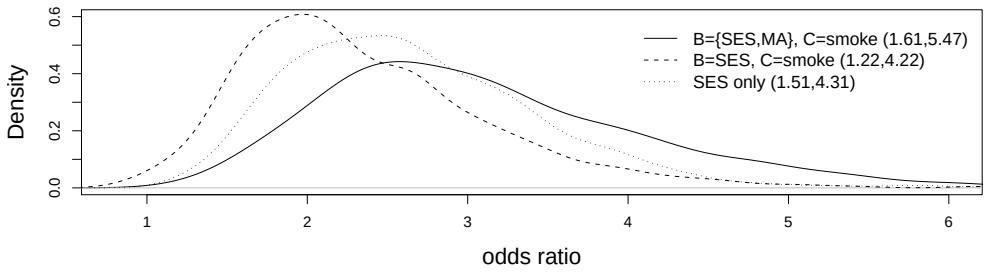


Figure 3: Kernel density plots for the posterior distribution of the PS adjusted OR estimates based on three different PS weights.

The results for the PS weighted and logistic regression estimates are similar across data sources. The frequentist estimates do not change whilst there is a change of 15 and 9 % for (CPCs,CPCn) and (CPM,EM) respectively. This indicates that some SB is being adjusted for probably due to differential participation rates in different strata of SES and maternal age.

While the (CPM, EM) estimates appear to increase, the (CPCs, CPCn) estimates vary across the data sources as shown in Table 5 and Figure 3. In particular the "B=SES, C=smoke" estimate is slightly lower, this may be

Adjusted odds ratios for H-H study

cas/con	Weighted		Un-weighted		
CPCs,CPCn			Freq LG		
Variables[a]	Med	CI	Variables[b]	Med	CI
$B = SES$	2.55	(1.51,4.31)	SES	2.62	(1.51,4.53)
$B = SES, C = smoke$	2.20	(1.22,4.22)	$SES, smoke$	2.62	(1.51,4.54)
$B = \{SES, MA\}, C = smoke$	2.94	(1.61,5.47)	$SES, smoke, MA$	2.58	(1.49,4.49)
CPM,EM					
$B = SES$	2.63	(1.56,4.61)			
$B = SES, C = smoke$	2.77	(1.58,5.13)			
$B = \{SES, MA\}, C = smoke$	2.86	(1.59,5.26)			

Table 5: *a*)Medians and 95% credible intervals for the odds ratios using (CPCs,CPCn),(CPM,EM) for different data sources, i.e. settings of the bias breaker B and the confounder C. *b*) Logistic regression estimates and 95% confidence intervals for the odds ratio adjusting for SES, then SES and $smoke$ and finally SES,$smoke$ and MA. In the logistic regression model, all three variables can be seen as confounder. SES is measured using the 1991 Carstairs tertiles, smoke refers to the mother's smoking status, MA refers to maternal age as categorised in Table 2.

due to some additional imprecision in the PS weights (see on-line supplementary material for details). The remaining PS weighted estimates resemble the (CPM, EM).

The PS adjusted estimates do not lead to changes in the epidemiological conclusions of the study, indicating that the results are robust to the choice of PS weights.

5 Discussion

In both studies we observe some sensitivity to different models and weights. However the estimates and the statistical significance of the results does not change in either study, suggesting that the original study estimates are robust and unlikely to be affected by substantial SB. In other cases, the statistical significance of the ORs could change. In particular, studies with highly variable ORs and large differences between the internal and external distributions of B will be the least robust.

Our PS based method provides a novel empirically and theoretically grounded approach to SB sensitivity analysis. The perturbations of the OR are caused by varying observed distributions of variables thought to be implicated in the selection. Further, PS weights to adjust for SB follow in a straightforward manner from the conditional independence structures that define this bias.[2] Our approach can be seen as complementing the bias parameter approach proposed by Greenland et al. [15,16] In particular, as the focus of our work is to adjust for SB, other forms of bias can still be modelled and taken into account via a bias parameter in the model for the exposure probability.

Inverse probability weighting (IPW) approaches are useful for adjusting for SB in particular in trials with dependent drop-out[17] where selection probabilities can typically be estimated. However, IPW often involves complex computations and smoothing procedures.[18] In contrast, our method is computationally simple, in particular when the weights are estimated using raw frequencies. It is thus worth experimenting with our method if external data can be found before moving on to more complex approaches. BPS can be easily extended if models are needed to estimate the PS weights, and to contexts outside case-control studies where SB is present.

When a suitable set of variables B to perform poststratification cannot be found, sensitivity analysis can be performed by proposing plausible distributions for B in the same spirit as [15] using prior knowledge to elicit and constrain these distributions.

Our method is computationally simple and fast especially with the independent multinomial-Dirichlet prior models we considered here as the posterior distributions can be derived analytically. In WinBUGs running these models took less than a minute on a standard PC.

Finally, BPS is conceptually simple and encourages us to think carefully about how case and control populations differ; What our target population is; What the sources of bias are and what variables we can use to assess the sensitivity of the results to these biases.

References

1. Mezei G, Kheifets L. Selection bias and its implications for case-control studies: a case study of magnetic field exposure and childhood leukaemia . *International Journal of Epidemiology* 2006; **35**:397–406.

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

3. *Non-ionizing radiation, Part 1: Static and Extremely low-frequency Electric and magnetic fields*, volume 80 of *IARC Monographs on the evaluation of Carcinogenic Risks to Humans*. IARC Press, Lyon, France , 2002.

4. Linet M, Hatch E, Kleinerman R, et al. Residential Exposure to Magnetic fields and acute lymphoblastic leukaemia in Children . *New Engl J Med* 1997; **337**(1):1–7.

5. Hatch E, Kleinerman R, Linet M, et al. Do confounding or selection factors of residential wire codings and magnetic fields distort findings of electromagnetic field studies? . *Epidemiology* 2000; **11**(2):189–198.

6. Ormond G, Nieuwenhuijsen M, Nelson P, et al. Endocrine Disruptors in the Workplace, Hair Spray, Folate Supplementation, and Risk of Hypospadias: Case-Control Study . *Environ Health Perspect* 2009; **117**(2):303–307.

7. Carstairs V, Morris R. *Deprivation and Health in Scotland*. Aberdeen University Press, Aberdeen , 1991.

8. Rotnitzky A, Robins J. Inverse probability weighted estimation in survival analysis . In *Encyclopedia of Biostatistics,* Armitage P, Colton T, editors, volume 4, 2619–2625. Wiley: New York , 2005.

9. Gelman A. Struggles with survey weighting and regression modelling . *Stat Sci* 2007; **22**(2):153–164.

10. Dawid A. P. Conditional Independence in Statistical Theory . *J R Stat Soc Series B Stat Methodol* 1979; **41**(1):1–31.

11. Hernan M, Hernandez-Diaz S, Robins J. A structural approach to selection bias . *Epidemiology* 2004; **15**(5):615–625.

12. Gelman A, Carlin B. Poststratification and weighting adjustments . In *Survey Nonresponse,* Groves R, Dillman D, Eltinge J, et al, editors, 289–302. New York: Wiley. , 2001.

13. Seaman S, Richardson S. Bayesian analysis of case-control studies with categorical covariates . *Biometrika* 2001; **88**(4):1073–1088.

14. Carmichael S, Shaw G, Laurent C, et al. Hypospadias and maternal exposures to cigarette smoke . *Paediatr Perinat Epidemiol* 2005; **19**(6):406–412.

15. Greenland S. Multiple-bias modelling for analysis of observational data . *J R Stat Soc Ser A Stat Soc* 2005; **168**(2):267–306.

16. Greenland S, Kheifets L. Leukemia Attributable to Residential Magnetic Fields: Results from Analyses Allowing for Study Biases . *Risk Anal* 2006; **26**(2).

17. Robins J, Rotnitzky A, Scharfstein D. *Statistical Models in Epidemiology: The Environment and Clinical Trials*, volume IMA 116, 1–92. NY: Springer-Verlag 1999.

18. Bang H, Robins J. Doubly Robust Estimation in Missing Data and Causal Inference Models . *Biometrics* 2004; **61**(4):962–972.

19. Lunn D, Thomas A, Best N, et al. WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility . *Statistics and Computing* 2000; **10**(4):325–337.