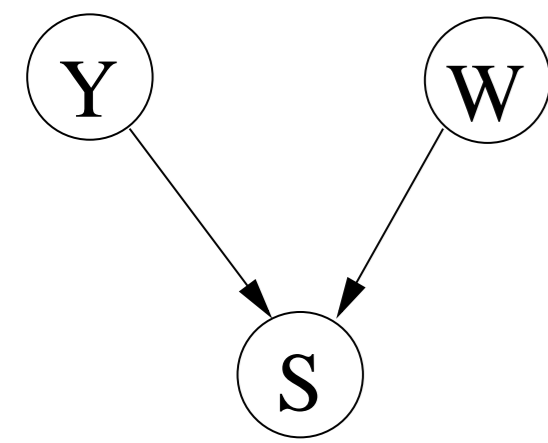


## Introduction

Selection bias in observational studies occurs when the association between an outcome and exposure of interest is induced or distorted by the nature of the individuals recruited into the study. Some examples of selection bias in Epidemiology are: **1. Biased case/control study base**, **2. Self-selection bias**, **3. Dependent drop-out**, **4. Berkson's bias**. In this poster we use Directed Acyclic Graphs<sup>1</sup> to model selection bias.

## DAGs

Denote by  $W$  an exposure of interest and by  $Y$  a disease under study. The DAG on the left encodes the conditional (in)dependences<sup>2</sup> below which show how conditioning on selection can result in a spurious association :



$$\begin{aligned} W &\perp\!\!\!\perp Y \\ W &\not\perp\!\!\!\perp Y|S \end{aligned}$$

In this case, the exposure and the disease are marginally independent - if we had data on the whole population we would infer that there is no association between  $W$  and  $Y$  - BUT conditional on selection  $W$ , is associated to  $Y$ .

Further, even when  $W$  and  $Y$  are truly associated, conditioning on selection can distort the strength of the association.

## Case control studies and odds ratios

Selection bias is introduced into the odds ratio as follows: Let  $Y$  and  $W$  be binary variables and let  $S$  take on value 1 for an individual selected into the study and 0 otherwise.

Odds ratio of interest

$$\psi = \frac{p(W=1|Y=1)p(W=0|Y=0)}{p(W=0|Y=1)p(W=1|Y=0)} = \frac{\pi^1 \times (1 - \pi^0)}{\pi^0 \times (1 - \pi^1)} \quad (1)$$

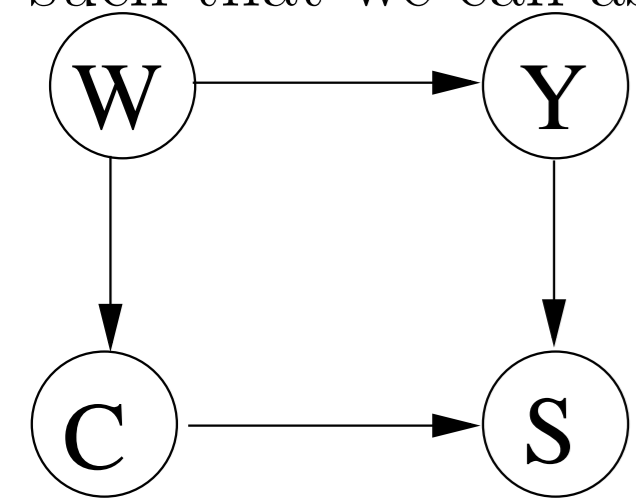
Observed odds ratio

$$\psi^o = \frac{p(W=1|Y=1, S=1)p(W=0|Y=0, S=1)}{p(W=1|Y=0, S=1)p(W=0|Y=1, S=1)} \quad (2)$$

The observed odds ratio (2) will not generally be the same as the "true" odds ratio (1) when  $S$  and  $W$  are associated.

## Possible solution

The problem is impossible to overcome without additional assumptions and/or information. In particular, the problem can be addressed if we can find an intermediate "bias breaking" variable  $C$  for which we have measurements for the whole eligible population, such that we can assume that



**Assumption A1**

$$\begin{aligned} Y &\perp\!\!\!\perp C|W \\ W &\perp\!\!\!\perp S|Y, C \end{aligned} \quad (3) \quad (4)$$

The DAGs on the left are two that encode (3) and (4). Note that these conditional independences are also encoded in the DAG with  $W \leftarrow Y$  and  $W \rightarrow C$ . For ease of interpretation we omit this DAG. The mathematics below relies only on the conditional independences and not on a particular DAG structure. Call this the **Bias breaking model**

Let  $C$  be stratified into  $k$  strata  $\mathcal{C}_i$   $i \in \{1, \dots, k\}$

Denote by  $p(W=1|Y=y, S=1, C \in \mathcal{C}_i) \equiv p(W=1|Y=y, C \in \mathcal{C}_i) = \pi_i^y$

Then:

$$\begin{aligned} \pi^y &= p(W=1|Y=y) = \sum_{i=1}^k p(W=1|Y=y, C \in \mathcal{C}_i)p(C \in \mathcal{C}_i|Y=y) \\ &= \sum_{i=1}^k \pi_i^y \times \alpha_i \end{aligned} \quad (5)$$

for  $y \in \{0, 1\}$  where  $\alpha_i = p(C \in \mathcal{C}_i|Y=y)$ . Thus, for  $y=1$ , (5) is the weighted sum of the observed exposed fraction of cases in each stratum of  $C$ . If  $\pi^y$  is not equal to  $p(W=1|Y=y, S=1)$ , the observed probability, then there could be selection bias due to the association between  $S$  and  $W$  via  $C$ .

## Estimates

We also assume that **A2: we can treat case and control selection as independent processes** and also, for the sake of simplicity that **A3 there is no control selection bias**: i.e. we can use the naive estimate  $p(W=1|Y=0, S=1)$  in (1). Then we can estimate  $\pi^1$  as follows:

$$\hat{\pi}^1 = \sum_{i=1}^k \hat{\pi}_i^1 \times \hat{\alpha}_i = \sum_{i=1}^k \left( \frac{m_i}{n_i} \times \frac{N_i}{N} \right) \quad (6)$$

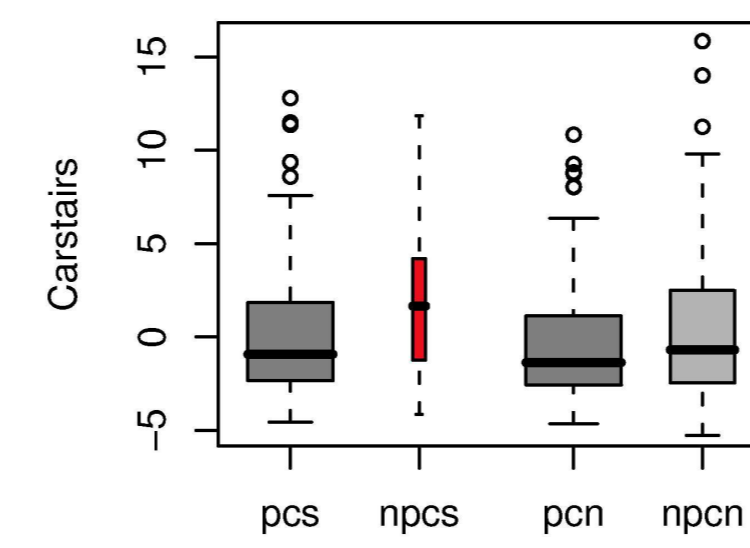
where  $m_i$  is the number of exposed participant cases in stratum  $\mathcal{C}_i$ ,  $n_i$  the number of participant cases in stratum  $\mathcal{C}_i$ ,  $N_i$  the total number of eligible cases in stratum  $\mathcal{C}_i$  and  $N$  is the total number of eligible cases. The naive estimate is given by  $\hat{\nu}^1$ .

$$\hat{\nu}^1 = \frac{m}{n} = \frac{\sum_{i=1}^k m_i}{\sum_{i=1}^k n_i}$$

where  $m$  is the number of exposed participant cases and  $n$  is the number of participant cases. A discrepancy between  $\hat{\pi}^1$  and  $\hat{\nu}^1$  could indicate selection bias via  $C$ .

## Application

We applied the above method to a case control study investigating the association between a congenital malformation and various lifestyle and employment related exposures. Cases were listed in a registry and were invited to participate through their GP, thus the wards of participants and non-participants were known. Controls were invited to participate via a postcard at their GP. Thus the wards of all respondent controls was known. Using the ward information, we were able to assign a **Carstairs** score<sup>3</sup>, a deprivation score to all cases as well as respondent controls.



To the left is a box-plot of the Carstairs scores for the cases and controls subdivided by participation: npcs are the non-participant cases, npcn are the non-participant controls. We assumed that the respondent controls were a representative sample of the eligible control population. However, there was concern about possible selection bias in the non-participant cases due to socio economic status (SES) as these had a substantially higher (i.e. more deprived) Carstairs score than the participants.

## Proposed model

We assume that the relationship between selection  $S$ , disease status  $Y$ , exposure  $W$  and SES indicated by Carstairs  $C$  is described by conditional independences (3) and (4) and thus the associated DAGs. This model says that selection bias occurs because the selection criteria are associated with the exposure through the (non)participants SES. To use the method developed above, we need to discretise the Carstairs score.

## Results

We applied the method presented above focusing on two exposures, smoking and maternal education. Below are the estimates for the adjusted and naive odds ratios:

- Naive odds ratio (the frequentist 95% confidence intervals): logistic regression in R - estimates given by (2)
- Adjusted odds ratio: divided Carstairs (range -5.3 to 15.9) score into 4 bins of minimum 1.5 width. An odds ratio was estimated for each possible ordered combination of the 4 bins using (6). The mean over all combinations was the adjusted estimate.

	Adjusted	Naive	Lower 95%	Upper 95%
Smoking	1.22	1.18	0.750	1.86
Maternal education	0.569	0.589	0.410	0.848

Although there was a difference in the naive and adjusted estimates, it was small.

## Simulation

To check the performance of the method in the presence of selection bias, we simulated a case control study similar to the one in the application but where the exposed were two times more likely to get the disease than the unexposed and the controls (not the cases) were subject to selection bias - the lower  $C$ , the less likely to respond. The table below left shows the simulation probabilities:

C	$p(C)$	$p(W C)$	$W$	$p(Y W)$
1	0.13	0.25	1	0.062
2	0.38	0.33	0	0.031
3	0.33	0.13		
4	0.17	0.062		
	$p(S Y=1)$	$p(S Y=0, C)$		
1	0.98	0.19		
2	-	0.38		
3	-	0.58		
4	-	0.58		

Below are the average adjusted odds ratio and their standard error, and the naive estimates and their standard error over 10 simulations of a base population of 100,000 - this gave approximately 900 participants.

Adj	se(Adj)	Nai	se(Nai)
2.08	$7.7 \times 10^{-4}$	2.43	0.023

The adjusted estimate is closer to the true odds ratio of 2 whereas the naive estimate consistently over-estimates the odds ratio.

## Conclusions

DAGs and conditional independences enabled us to clearly state our beliefs and assumptions about how selection bias occur and further how to formulate a model to overcome selection bias.

Using the bias breaking model to reduce selection bias worked well in the simulation study with the adjusted estimate coming closer to the true value. Further the systematic divergence between the adjusted and naive estimates clearly indicated the presence of selection bias. Note that the response rate of the cases was 98% in the simulation - a lower response rate could blur the distinction between the adjusted and naive estimates.

Comparing these results to those of the application we can conclude that if there is any selection bias via SES then it is very weak. Importantly, it does not change the conclusions reached by the study.

## Further work

If we believe that the data generating process could be modelled by a bias breaking model but that data are not available on a bias breaking variable, it could be possible to replace  $p(C|Y) = \alpha$  with a prior representing our beliefs about how bias is being introduced. We need to relate the above to similar approaches in the survey sampling literature.

## References and Acknowledgments

- Hernan, M.A., Hernandez-Diaz, S., Robins, J.M. *A structural approach to selection bias* Epidemiology, 15(5), 2004.
  - Dawid, A.P. *Conditional Independence in Statistical Theory* JRSS B, 41(5), 1979
  - Carstairs, V. and Morris, R. *Deprivation and Health in Scotland*, Aberdeen University Press, 1991.
- Thanks to the ESRC for funding and to Paul Nelson.