Simple sensitivity analysis for control selection bias

Louisa H. Smith & Tyler J. VanderWeele

To appear in Epidemiology. Pre-peer-review version.

Case-control studies allow for efficient sampling schemes but are subject to bias when controls fail to represent the exposure distribution in the population from which the cases were sampled. Identifying this population, known as the study base, is often a challenge, and controls may be chosen out of convenience or to avoid other types of bias, such as exposure misclassification.[1] On the other hand, it may be straightforward to completely ascertain or randomly sample cases, as they may be enumerated in registries, hospital records, or other sampling frames.

When inappropriate control selection is suspected to have occurred, it can be informative to conduct a sensitivity analysis to investigate the possible extent of the resulting bias. In this letter we show that a recently developed framework for simple sensitivity analysis[2–4] can be extended to this situation. We demonstrate with an example, and we provide a more detailed derivation in the Supplement.

MacMahon et al. conducted a case-control study of pancreatic cancer patients, whom they compared to controls who were patients of the same physicians as the cases, but who had different illnesses.[5] After adjusting for age, sex, and cigarette smoking, they found an odds ratio of 2.7 (95% CI 1.6, 4.7) comparing drinkers of at least 3 cups per day to non-coffee-drinkers.

However, soon after the study was published, multiple possible sources of bias were described.[6] In particular, many of the control patients had gastrointestinal disorders, which the investigators failed to account for. If the controls drank less coffee than the general source population due to their illnesses, selection bias would result, exaggerating the association between coffee and pancreatic cancer.

To quantify the possible size of this bias, consider the ratio of the observable odds ratio from case-control data (OR_{obs}) to the odds ratio that would have been estimated had the entire study base been sampled (OR_{true}). For simplicity, assume that any bias from the case-control study is due to poor control selection.

It is possible to derive a bound similar to that in Smith & VanderWeele 2019,[4] but with different definitions for the parameters resulting from the different causal structure (Figure)

and estimand of interest. Specifically, if we assume that selection ($S = 1$) of cases ($Y = 1$) is independent of exposure status ($A \in \{0, 1\}$) (possibly conditional on measured covariates C), but that control ($Y = 0$) selection is not independent of exposure without additionally conditioning on unmeasured factor(s) U, then:

$$\mathrm{OR}_{\mathrm{obs}}/\mathrm{OR}_{\mathrm{true}} \leq \left\{ \frac{\mathrm{RR}_{UA_1} \times \mathrm{RR}_{S_0U}}{\mathrm{RR}_{UA_1} + \mathrm{RR}_{S_0U} - 1} \right\} \times \left\{ \frac{\mathrm{RR}_{UA_0} \times \mathrm{RR}_{S_1U}}{\mathrm{RR}_{UA_0} + \mathrm{RR}_{S_1U} - 1} \right\}$$

where

$$\begin{aligned}
\mathrm{RR}_{UA_1} &= \frac{\max_u P(A = 1|Y = 0, u, c)}{\min_u P(A = 1|Y = 0, u, c)} \\
\mathrm{RR}_{UA_0} &= \frac{\max_u P(A = 0|Y = 0, u, c)}{\min_u P(A = 0|Y = 0, u, c)} \\
\mathrm{RR}_{S_1U} &= \max_u \frac{P(U = u|Y = 0, S = 1, c)}{P(U = u|Y = 0, S = 0, c)} \\
\mathrm{RR}_{S_0U} &= \max_u \frac{P(U = u|Y = 0, S = 0, c)}{P(U = u|Y = 0, S = 1, c)} \ .
\end{aligned}$$

To understand these parameters, suppose that U represents a binary indicator of gastrointestinal illness that affects coffee-drinking and also makes hospital visits (and therefore selection as a control) more likely. With respect to the example, RR_{UA_1} describes the increased probability of drinking ≥ 3 cups of coffee per day in eligible controls *without* gastrointestinal disorders compared to those *with*, RR_{UA_0} is the increased probability of no coffee drinking in eligible controls *with* gastrointestinal disorders compared to those *without*, RR_{S_1U} is the increased probability of GI disorders in controls who *were* selected for the study compared to those who were *not*, and RR_{S_0U} is the increased probability of a health GI system in controls who were *not* selected for the study compared to those who *were*.

We could generate plausible values for these parameters to "correct" for selection bias, or we couldWe could propose various values for these parameters to "correct" for, or bound, selection bias. For example, suppose that among eligible controls with gastrointestinal disorders, only 5% drink at least 3 cups of coffee daily. However, among those with healthy gastrointestinal tracts, 30% drink that amount. Then $\mathrm{RR}_{UA_1} = 0.3/0.05 = 6$, and $\mathrm{RR}_{UA_0} = 0.95/0.7 = 1.36$. Next suppose that among selected controls, the prevalence of gastrointestinal disorders is 0.45, but among non-selected eligible controls it is 0.1. Assuming for the purposes of the example that gastrointestinal disorders is binary, then $\mathrm{RR}_{S_1U} = 0.45/0.1 = 4.5$ and $\mathrm{RR}_{S_0U} = 0.9/0.55 = 1.64$, then using these values in the formula for bound above we would obtain 1.87. Thus we would

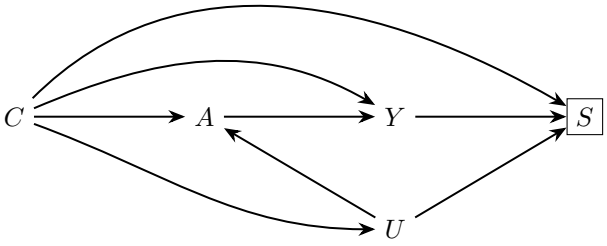

Figure 1: A directed acyclic graph describing a causal structure that could lead to selection bias. In the example in the text, A is coffee consumption, Y pancreatic cancer, S selection into the study, C measured covariates, and U gastrointestinal disorders.

have$\text{OR}_{\text{true}} \geq 2.7/1.87 = 1.44$, where 2.7 was the observed odds ratio and 1.87 the bound constructed from the proposed parameters. We could repeat this exercise with a range of other values or allow for a more complex unmeasured factor (e.g., severe gastrointestinal disorder, mild discomfort, health gastrointestinal tract), as well as repeat with the lower bound of the confidence interval.

As with the related sensitivity analysis bounds, it is also possible to calculate a single value describing the strength the parameters would have to have to have in order for the observed odds ratios be entirely due to selection bias. However, this "selection bias E-value," which has the same formula as in Smith & VanderWeele 2019,[4] $\sqrt{\text{OR}_{\text{obs}}} + \sqrt{\text{OR}_{\text{obs}} - \sqrt{\text{OR}_{\text{obs}}}}$ but which instead refers to the value that jointly minimizes the above parameters, is not as easily interpretable due to the fact that these parameters do not vary independent of one another.

To make this type of sensitivity analysis easy to perform, we have created an online calculator available at www.selection-bias.com.

REFERENCES

1. Wacholder S, McLaughlin JK, Silverman DT, et al. Selection of controls in case-control studies: I. Principles. *Am J Epidemiol.* 1992;135:1019–1028.

2. Ding P, VanderWeele TJ. Sensitivity analysis without assumptions. *Epidemiology.* 2016;27:368–377.

3. VanderWeele TJ, Ding P. Sensitivity analysis in observational research: Introducing the e-value. *Ann Intern Med.* 2017;167:268–275.

4. Smith LH, VanderWeele TJ. Bounding bias due to selection. *Epidemiology*. 2019;30:509–516.

5. MacMahon B, Yen S, Trichopoulos D, et al. Coffee and cancer of the pancreas. *New Engl J Med*. 1981;304:630–633.

6. Feinstein AR, Horwitz RI, Spitzer WO. Coffee and pancreatic cancer: The problems of etiologic science and epidemiologic case-control research. *JAMA*. 1981;246:957.

Supplement for *Simple sensitivity analysis for control selection bias*

In a case-control study we can estimate the odds ratio, conditional on covariates C,

$$\frac{\Pr(Y=1 \mid A=1, S=1, c)}{\Pr(Y=0 \mid A=1, S=1, c)} \Big/ \frac{\Pr(Y=1 \mid A=0, S=1, c)}{\Pr(Y=0 \mid A=0, S=1, c)}$$

where Y indicates case vs. control status, A is the binary exposure of interest, and S is an indicator of selection into the case-control study.

This quantity can be used to estimate the population odds ratio

$$\frac{\Pr(Y=1 \mid A=1, c)}{\Pr(Y=0 \mid A=1, c)} \Big/ \frac{\Pr(Y=1 \mid A=0, c)}{\Pr(Y=0 \mid A=0, c)}$$

without bias, as long as $\Pr(A=1 \mid Y=0, S=1, c) = \Pr(A=1 \mid Y=0, c)$ and $\Pr(A=1 \mid Y=1, S=1, c) = \Pr(A=1 \mid Y=1, c)$. In other words, selection of both cases and controls must be independent of exposure.

Although it may be straightforward to randomly sample the cases with respect to the distribution of their exposure, often because the cases can be fully enumerated, control selection is usually more difficult. When the sampled controls do not represent the exposure distribution in the source population, selection bias can result.

To quantify the possible size of this bias, consider the ratio of the observable odds ratio from case-control data to the odds ratio that would have been estimated had the entire cohort been sampled. Assume then that any bias from the case-control study is due to this selection.

We therefore have:

$$\text{bias} = \left\{ \frac{\Pr(Y=1 \mid A=1, S=1, c)}{\Pr(Y=0 \mid A=1, S=1, c)} \Big/ \frac{\Pr(Y=1 \mid A=0, S=1, c)}{\Pr(Y=0 \mid A=0, S=1, c)} \right\} \Big/ \left\{ \frac{\Pr(Y=1 \mid A=1, c)}{\Pr(Y=0 \mid A=1, c)} \Big/ \frac{\Pr(Y=1 \mid A=0, c)}{\Pr(Y=0 \mid A=0, c)} \right\}.$$

We can rewrite each odds ratio in terms of the probability of the exposure:

$$\text{bias} = \left\{ \frac{\Pr(A=1 \mid Y=1, S=1, c)}{\Pr(A=0 \mid Y=1, S=1, c)} \Big/ \frac{\Pr(A=1 \mid Y=0, S=1, c)}{\Pr(A=0 \mid Y=0, S=1, c)} \right\} \Big/ \left\{ \frac{\Pr(A=1 \mid Y=1, c)}{\Pr(A=0 \mid Y=1, c)} \Big/ \frac{\Pr(A=1 \mid Y=0, c)}{\Pr(A=0 \mid Y=0, c)} \right\}.$$

Now assume that the cases have been properly sampled independently of exposure status, such that $A \perp\!\!\!\perp S \mid Y = 1$, but that the independence does not hold for $Y = 0$:

$$\text{bias } = \frac{\Pr(A = 1 \mid Y = 0, c)}{\Pr(A = 0 \mid Y = 0, c)} \Big/ \frac{\Pr(A = 1 \mid Y = 0, S = 1, c)}{\Pr(A = 0 \mid Y = 0, S = 1, c)} .$$

Following the logic in Smith & Vanderweele 2019,[1] we see that

$$\begin{aligned}
\text{bias } &\leq \frac{\max_s \Pr(A = 1 \mid Y = 0, S = s, c)}{\min_s \Pr(A = 0 \mid Y = 0, S = s, c)} \Big/ \frac{\Pr(A = 1 \mid Y = 0, S = 1, c)}{\Pr(A = 0 \mid Y = 0, S = 1, c)} \\
&\leq \frac{\Pr(A = 1 \mid Y = 0, S = 0, c)}{\Pr(A = 0 \mid Y = 0, S = 0, c)} \Big/ \frac{\Pr(A = 1 \mid Y = 0, S = 1, c)}{\Pr(A = 0 \mid Y = 0, S = 1, c)} \\
&= \frac{\Pr(A = 1 \mid Y = 0, S = 0, c)}{\Pr(A = 1 \mid Y = 0, S = 1, c)} \Big/ \frac{\Pr(A = 0 \mid Y = 0, S = 0, c)}{\Pr(A = 0 \mid Y = 0, S = 1, c)} .
\end{aligned}$$

Suppose there exists some U such that $A \perp\!\!\!\perp S \mid Y = 0, C, U$. For notational simplicity we will assume discrete U. Then we can write, by Lemma A.3 in Ding & VanderWeele 2016:[2]

$$\begin{aligned}
\text{bias } &\leq \left\{ \frac{\sum_u \Pr(A = 1 \mid Y = 0, S = 0, c, u) \Pr(U = u \mid Y = 0, S = 0, c)}{\sum_u \Pr(A = 1 \mid Y = 0, S = 1, c, u) \Pr(U = u \mid Y = 0, S = 1, c)} \right\} \Big/ \\
&\qquad \left\{ \frac{\sum_u \Pr(A = 0 \mid Y = 0, S = 0, c, u) \Pr(U = u \mid Y = 0, S = 0, c)}{\sum_u \Pr(A = 0 \mid Y = 0, S = 1, c, u) \Pr(U = u \mid Y = 0, S = 1, c)} \right\} \\
&= \left\{ \frac{\sum_u \Pr(A = 1 \mid Y = 0, c, u) \Pr(U = u \mid Y = 0, S = 0, c)}{\sum_u \Pr(A = 1 \mid Y = 0, c, u) \Pr(U = u \mid Y = 0, S = 1, c)} \right\} \Big/ \\
&\qquad \left\{ \frac{\sum_u \Pr(A = 0 \mid Y = 0, c, u) \Pr(U = u \mid Y = 0, S = 0, c)}{\sum_u \Pr(A = 0 \mid Y = 0, c, u) \Pr(U = u \mid Y = 0, S = 1, c)} \right\} \\
&\leq \left\{ \frac{\text{RR}_{UA_1} \times \text{RR}_{S_0U}}{\text{RR}_{UA_1} + \text{RR}_{S_0U} - 1} \right\} \times \left\{ \frac{\text{RR}_{UA_0} \times \text{RR}_{S_1U}}{\text{RR}_{UA_0} + \text{RR}_{S_1U} - 1} \right\}
\end{aligned}$$

where

$$\begin{aligned}
\text{RR}_{UA_1} &= \frac{\max_u \Pr(A = 1 | Y = 0, u, c)}{\min_u \Pr(A = 1 | Y = 0, u, c)} \\
\text{RR}_{UA_0} &= \frac{\max_u \Pr(A = 0 | Y = 0, u, c)}{\min_u \Pr(A = 0 | Y = 0, u, c)} \\
\text{RR}_{S_1U} &= \max_u \frac{\Pr(U = u | Y = 0, S = 1, c)}{\Pr(U = u | Y = 0, S = 0, c)} \\
\text{RR}_{S_0U} &= \max_u \frac{\Pr(U = u | Y = 0, S = 0, c)}{\Pr(U = u | Y = 0, S = 1, c)} .
\end{aligned}$$

REFERENCES

1. Smith LH, VanderWeele TJ. Bounding bias due to selection. *Epidemiology.* 2019;30:509–516.

2. Ding P, VanderWeele TJ. Sharp sensitivity bounds for mediation under unmeasured mediator-outcome confounding. *Biometrika.* 2016;103:483–490.