ACM SIGKDD Explorations Newsletter. 2004;6:1–6.
3. Naimi AI, Balzer LB. Stacked generalization: an introduction to super learning. *Eur J Epidemiol*. 2018;33:459–464.
4. *SuperLearner: Super Learner Prediction*. [computer program]. Version R package version 2.0-2.4. Available at: https://CRAN.R-project.org/package=SuperLearner2018. Accessed 1 October 2019.
5. Batista G, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*. 2004;6:20–29.
6. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artifi Intell Res*. 2002;16:321–357.
7. Kuhn M. Building predictive models in R using the caret package. *J Stat Soft*. 2008;28:1–26.
8. *pROC: an open-source package for R and S+ to analyze and compare ROC curves* [computer program]. *BMC Bioinformatics*. 2011;12:77.

Simple Sensitivity Analysis for Control Selection Bias

To the Editor:

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 eAppendix; http://links.lww.com/EDE/B670.

MacMahon et al[5] conducted a case–control study of pancreatic cancer patients whom they compared with controls who were patients of the same physicians as the cases but who had different illnesses. After adjusting for age, cigarette smoking, and sex, they found an odds ratio of 2.7 (95% confidence interval = 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 and VanderWeele[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:

$$OR_{\text{obs}} / OR_{\text{true}} \leq \left\{\frac{RR_{UA_1} \times RR_{S_0U}}{RR_{UA_1} + RR_{S_0U} - 1}\right\} \times \left\{\frac{RR_{UA_0} \times RR_{S_1U}}{RR_{UA_0} + RR_{S_1U} - 1}\right\}$$

where

$$RR_{UA_1} = \frac{\max_u \Pr(A = 1 \mid Y = 0, u, c)}{\min_u \Pr(A = 1 \mid Y = 0, u, c)}$$

$$RR_{UA_0} = \frac{\max_u \Pr(A = 0 \mid Y = 0, u, c)}{\min_u \Pr(A = 0 \mid Y = 0, u, c)}$$

$$RR_{S_1U} = \max_u \frac{\Pr(U = u \mid Y = 0, S = 1, c)}{\Pr(U = u \mid Y = 0, S = 0, c)}$$

$$RR_{S_0U} = \max_u \frac{\Pr(U = u \mid Y = 0, S = 0, c)}{\Pr(U = u \mid Y = 0, S = 1, c)}.$$

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 with those with gastrointestinal disorders, RR_{UA_0} is the increased probability of no coffee drinking in eligible controls with gastrointestinal disorders compared with those without gastrointestinal disorders, RR_{S_1U} is the increased probability of gastrointestinal disorders in controls who were selected for the study compared with those who were not, and RR_{S_0U} is the increased probability of a healthy GI system in controls who were *not* selected for the study compared with those who *were* selected.

We 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 $RR_{UA_1} = 0.3/0.05 = 6$

This work was supported by grant R01CA222147 from the National Institutes of Health (T.V.W.).
The authors report no conflicts of interest.
SDC Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com).
The analysis can be replicated at the website linked in the manuscript, http://selection-bias.louisahsmith.com.

ISSN: 1044-3983/2020/3105-e44
DOI: 10.1097/EDE.0000000000001207

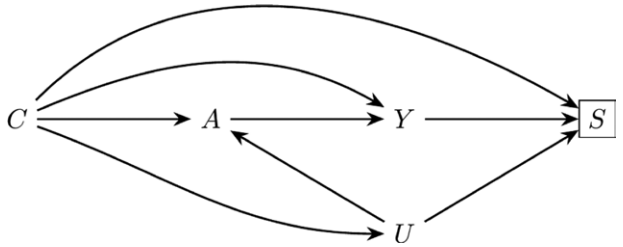

FIGURE. A directed acyclic graph describing a causal structure that could lead to selection bias. In the example in the text, *A* is coffee consumption, *Y* is pancreatic cancer, *S* is selection into the study, *C* is measured covariates, and *U* is gastrointestinal disorders.

and $RR_{UA_0} = 0.95/0.7 = 1.36$. Next suppose that among selected controls, the prevalence of gastrointestinal disorders is 0.45, but among nonselected eligible controls, it is 0.1. Assuming for the purposes of the example that gastrointestinal disorders is binary, then $RR_{S_1U} = 0.45/0.1 = 4.5$ and $RR_{S_0U} = 0.9/0.55 = 1.64$, then using these values in the formula for bound above, we would obtain 1.87. Thus we would have $OR_{\text{true}} \geq 2.7 / 1.87 = 1.44$, where 2.7 was the observed odds ratio and 1.87 the bound constructed from the proposed parameters. In other words, if we assume that eligible study participants outside the hospital had a 10% prevalence of gastrointestinal disorders and those without such disorders were 6 times more likely to be heavy coffee drinkers than those with such disorders, then had they been included in the sampling frame, 1.44 is a lower bound for the estimate of the odds ratio relating coffee drinking and pancreatic cancer, conditional on sex, age, and smoking status.

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, healthy gastrointestinal tract), as well as repeat with the lower bound of the confidence interval.

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

Louisa H. Smith
Department of Epidemiology
Harvard T.H. Chan School of Public Health
Boston, MA
Louisa_h_smith@g.harvard.edu

Tyler J. VanderWeele
Departments of Epidemiology
and Biostatistics
Harvard T.H. Chan School of Public Health
Boston, MA

REFERENCES

1. Wacholder S, McLaughlin JK, Silverman DT, Mandel JS. 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;167:268–274.
4. Smith LH, VanderWeele TJ. Bounding bias due to selection. *Epidemiology*. 2019;30:509–516.
5. MacMahon B, Yen S, Trichopoulos D, Warren K, Nardi G. Coffee and cancer of the pancreas. *N Engl J Med*. 1981;304:630–633.
6. Feinstein AR, Horwitz RI, Spitzer WO, Battista RN. Coffee and pancreatic cancer. The problems of etiologic science and epidemiologic case-control research. *JAMA*. 1981;246:957–961.