Published in final edited form as:
Epidemiology. 2012 September ; 23(5): 772–773. doi:10.1097/EDE.0b013e3182628365.

Using propensity scores to reduce case-control selection bias

Matthew C. Walsh,
Department of Population Health Sciences, School of Medicine and Public Health, University of Wisconsin - Madison, Madison, WI, walsh2@wisc.edu

Amy Trentham-Dietz,
Department of Population Health Sciences, & UW Paul P. Carbone Cancer Center, School of Medicine and Public Health, University of Wisconsin-Madison, Madison, WI

Polly A. Newcomb,
Cancer Prevention Program, Fred Hutchinson Cancer Research Center, Seattle, WA

Ronald Gangnon, and
Department of Population Health Sciences & Department of Biostatistics & Medical Informatics School of Medicine and Public Health University of Wisconsin & Madison, Madison, WI

Mari Palta
Department of Population Health Sciences & Department of Biostatistics & Medical Informatics School of Medicine and Public Health University of Wisconsin – Madison, Madison, WI

To the Editor:

When sampling frames differ for cases and controls in a population based study, there are several analytical options to increase study validity. One is to exclude cases not found in the sampling frame for controls. Another, when the case sampling frame is nearly complete, is to link cases to the sampling frame for controls in order to obtain the information needed to model coverage propensity scores.[1] Propensity scores have been used to address selection bias due to non-response and treatment-selection,[2, 3] but we have found no studies that used propensity scores to address selection bias due to inadequate sampling frame coverage. We describe such an application here.

Our analyses use data from the Wisconsin Women's Health Study. Cases were women with incident invasive breast cancer reported to Wisconsin's mandatory cancer registry. Controls were identified using a master file of licensed drivers who consented to have their private information released by the Wisconsin Department of Transportation. An estimated 68% of Wisconsin women with valid driver's licenses provided this consent in 2008.[4] Case eligibility was limited to women with a Wisconsin driver's license (verified by self-report).

We calculated odds ratio (OR) estimates of breast cancer risk using 3 established methods to correct selection bias, and compared these with published estimates.[5, 6] The 3 methods were (1) excluding cases not in the control sampling frame, (2) applying propensity scores as weights, and (3) using quintiles of propensity scores to stratify analyses. We calculated standard errors of OR estimates to compare their precision. Additional details, including study inclusion criteria, are described elsewhere.[4]

The 2,988 breast cancer cases were linked to the driver's license file to determine whether cases matched a record from the master-file of drivers (sampling frame for controls).[4, 7] Propensity scores for inclusion in the sampling frame were calculated using logistic regression of 24 potential confounding factors ascertained through interviews with participants.

The Table shows results of applying the 3 correction methods to a re-analysis of antidepressant medications and breast cancer risk.[6] Of the 3 methods, stratifying by quintile of propensity score produced odds ratio[3] with the best precision. Weighting, stratification and exclusion produced odds ratio[3] with a similar degree of bias correction. However, the odds ratio3 changed only slightly after applying any of the correction methods. Similar results were found in another re-analysis of breast cancer risk, that one in relation to reproductive factors.[4]

Historically, the standard procedure in epidemiologic case-control studies has been to exclude cases that could not have been approached to serve as controls.[8] This procedure assumes that factors that predict inclusion in the control sampling frame are similar for cases and potential controls. The two propensity-score-based methods implemented here also require this assumption and in addition, require correct model specification for propensity score estimation. The advantage of the propensity score methods is that they preserve the study base by allowing all cases to be used.

Selection bias caused by incomplete sampling-frame coverage is frequently overlooked. In this study we linked the case and control sampling frames to demonstrate the use of propensity scores to adjust for selection bias due to incomplete study-base coverage of the sampling frames.

Acknowledgments

We gratefully acknowledge study assistance provided by Leonelo Bautista, Nora Cate Schaeffer, Javier Nieto, Paul Peppard, John Hampton, Julie McGregor, and Laura Stephenson.

Funding: supported by the National Institutes of Health grants CA47147, CA67264 and CD000712

REFERENCES

1. Rosenbaum P, Rubin D. The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika. 1983; 70(1):41–55.
2. Austin PC. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. Med Decis Making. 2009; 29(6): 661–677. [PubMed: 19684288]
3. Mansson R, Joffe MM, Sun W, Hennessy S. On the estimation and use of propensity scores in case-control and case-cohort studies. Am J Epidemiol. 2007; 166(3):332–339. [PubMed: 17504780]
4. Walsh, M. Improving case-control study validity by addressing driver's license sampling frame limitations. Madison, WI: The University of Wisconsin - Madison; 2010.
5. Nichols HB, Trentham-Dietz A, Sprague BL, Hampton JM, Titus-Ernstoff L, Newcomb PA. Effects of birth order and maternal age on breast cancer risk: modification by whether women had been breast-fed. Epidemiology. 2008; 19(3):417–423. [PubMed: 18379425]
6. Wernli KJ, Hampton JM, Trentham-Dietz A, Newcomb PA. Antidepressant medication use and breast cancer risk. Pharmacoepidemiol Drug Saf. 2009; 18(4):284–290. [PubMed: 19226540]
7. Walsh M, Trentham-Dietz A, Gangnon R, Nieto F, Newcomb P, Palta M. Selection bias in population based cancer case-control studies due to incomplete sampling frame coverage. Cancer Epidemiology, Biomarkers & Prevention April 6, 2012.

8. Wacholder S, McLaughlin JK, Silverman DT, Mandel JS. Selection of controls in case-control studies. I Principles. Am J Epidemiol. 1992; 135(9):1019–1028. [PubMed: 1595688]

Table

Previously published[6] and selection-bias-corrected odds ratios[a] and 95% confidence intervals (CIs)[b] for breast cancer risk in relation to use of antidepressant medications, Wisconsin, 2004–2008

	Ever Use				Antidepressant Use Former Use				Current Use			
		(95% CI)				(95% CI)				(95% CI)		
	OR	Standard	Bootstrap	SE of OR	OR	Standard	Bootstrap	SE of OR	OR	Standard	Bootstrap	SE of OR
Published Odds Ratios	0.89	(0.78–1.01)		0.0651	0.82	(0.66–1.02)		0.1107	0.92	(0.80–1.07)		0.0745
Corrected Odds Ratios Exclusin[c]	0.86	(0.74–0.99)		0.0732	0.75	(0.58–0.96)		0.1261	0.90	(0.77–1.06)		0.0837
Coverage Propensity Score as Weights	0.86	(0.75–0.98)	(0.73–1.00)	0.0781	0.77	(0.62–0.96)	(0.59–1.01)	0.1320	0.90	(0.77–1.04)	(0.75–1.06)	0.0893
Quintiles of Propensity Scores	0.87	(0.75–0.99)	(0.76–1.00)	0.0702	0.78	(0.63–0.98)	(0.63–1.01)	0.1165	0.91	(0.78–1.05)	(0.79–1.07)	0.0798

[a] Unadjusted Odds Ratios from Table 2 in Wernli et al [6]; adjusted for age, year of interview, parity, age at first live birth, family history of cancer, body mass index, menopausal status, age at menopause, mammography, and type of hormone use. Reference category for all analyses was report of having never used antidepressants

[b] Robust variance estimates used for all confidence intervals; nonparametric bootstrap method used to obtain confidence intervals for coverage propensity score and quintiles of coverage correction methods

[c] Excluded cases not in the sampling frame were used to select controls