Commentary

A Double Robust Approach to Causal Effects in Case-Control Studies

Sherri Rose* and Mark van der Laan

* Correspondence to Dr. Sherri Rose, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115 (e-mail: rose@hcp.med.harvard.edu).

Initially submitted December 9, 2011; accepted for publication April 23, 2013.

In a recent issue of the *Journal*, VanderWeele and Vansteelandt (*Am J Epidemiol.* 2011;174(10):1197–1203) discussed an inverse probability weighting method for case-control studies that could be used to estimate an additive interaction effect, referred to as the “relative excess risk due to interaction.” In this article, we reinforce the well-known disadvantages of inverse probability weighting and comment on the desirability of the described parameter. Further, we review an existing double robust estimator not considered by VanderWeele and Vansteelandt, the case-control-weighted targeted maximum likelihood estimator, which has improved properties in comparison with a previously described inverse-probability-weighted estimator. This targeted maximum likelihood estimator can be used to target various parameters of interest, and its implementation has been described previously for the risk difference, relative risk, and odds ratio.

case-control studies; causality; epidemiologic methods; estimation techniques

Abbreviations: IPW, inverse probability weighting; RERI, relative excess risk due to interaction; TMLE, targeted maximum likelihood estimator.

Editor’s note: An invited commentary on this article appears on page 670, and the authors’ response appears on page 672.

Case-control study designs, where researchers sample from the target population conditional on the outcome (often a disease), are popular in epidemiology and medicine. The case-control design allows investigators to study outcomes whose observation would otherwise require extensive sampling and lengthy follow-up time, and it is frequently used to study rare diseases. Case-control sampling designs are also sometimes referred to as biased sampling designs in the statistics literature. This terminology is used because the resulting sample is biased with respect to the proportion of diseased individuals compared with the target population (1). For example, in a case-control study where the researchers sample 1 control (nondiseased subject) for every 1 case (diseased subject), the “prevalence” of disease in the sample is 50%. However, the prevalence of disease in the target population is likely to be much smaller, often less than 5% (1, 2).

Methods for case-control data have mostly focused on parametric logistic regression, from which one can estimate the odds ratio conditional on the covariates in the regression (this parameter differs from a marginal population-level odds ratio that may be of interest, particularly in policy-related research and causal inference questions). This conditional parameter depends on correct specification of the logistic regression, which is uncertain in practice. It is also redefined each time different covariates are considered in the regression. An odds ratio estimated from a 2×2 table of the outcome and exposure in an unmatched case-control study targets a marginal statistical parameter. However, this estimator does not adjust for covariates, and therefore it does not attempt to control confounding, which is a strong concern in observational studies. Additional methods for marginal parameters in case-control studies have been lacking compared with those available for general observational studies, and they include the “approximately valid” inverse probability weighting (IPW) method for case-control data (3) and the double robust case-control-weighted targeted maximum likelihood estimator (TMLE) (4–7).

The use of IPW (8, 9) for cross-sectional or cohort studies has a noteworthy recent history in the epidemiologic and causal inference literature (10–16). While its ease of implementation has contributed to its use, IPW methods have substantial and well-known drawbacks (4, 17–19). IPW estimators will be asymptotically unbiased if the probability of the exposure given covariates, used to define the weights, is consistently estimated (i.e., converges to the truth). Thus, correctly estimating this probability distribution is crucial. Greater issues arise when examining efficiency, as IPW estimators can make no claims regarding asymptotic efficiency and also have demonstrated efficiency performance problems in finite samples. These finite sample problems can be understood intuitively. One is weighting by the inverse probability of being exposed given covariates; when this probability is small, the weights will be very large, leading to sizeable standard errors and broad confidence intervals.

The efficiency and bias of IPW estimators is of particular concern when few or no subjects within particular covariate strata are exposed (or unexposed), which is also referred to as a positivity violation (or a violation of the experimental treatment assignment assumption). These positivity violations can be theoretical or empirical. Theoretical violations occur when it is not possible for subjects with given covariate values to be exposed (or unexposed). Empirical positivity violations appear when subjects with certain covariate values are not exposed (or unexposed) because of chance in the sample, possibly due to sample size. However, even with larger samples, empirical positivity violations can still occur, particularly when covariates are continuous. Imagine, for example, that there are empirical positivity violations in a study of the effect of a particular toxin on cancer with 1 binary covariate, sex. It would result in very few women being exposed to the toxin, leading the probability of exposure among women to be very small. In this case, women in the study would have very large inverse probability weights, and the IPW estimator will depend heavily on a few extreme values leading to finite sample bias. Positivity violations and near violations should be addressed when using any estimator, and diagnostic approaches are available to assess the bias induced by such violations (20). It is important to note that finite sample efficiency concerns with IPW estimators go beyond positivity violations. IPW estimators are not substitution estimators (or “plug-in” estimators); thus, they ignore global constraints and negatively impact finite sample efficiency.

The “approximately valid” IPW estimator for the odds ratio in case-control data (3) has not been widely adopted. In a new paper, VanderWeele and Vansteelandt (21) recently described an IPW method for case-control data and a particular target parameter, the “relative excess risk due to interaction” (RERI), which they approximate using odds ratios. Just as in the “approximately valid” IPW method originally presented by Robins for case-control data (3), they require that the outcome is rare and that the probability of the exposure given covariates is fitted among controls only. While VanderWeele and Vansteelandt note their forthcoming work in developing double robust estimators for unmatched case-control data, they do not discuss the existing literature on double robust case-control-weighted TMLEs (4–7). They also make strong claims about the desirability of an interaction parameter on the additive scale, asserting that these measures

are considered the most important parameters for public health applications (21). We counter that such a generalization does not encompass the breadth of research questions encountered in public health, and one might instead choose to focus on estimators that are more flexible and capable of estimating a broader range of parameters. This commentary was written to draw attention to causal methods for case-control studies overlooked by VanderWeele and Vansteelandt, notably the double robust case-control-weighted TMLE, which can currently estimate the causal risk difference, relative risk, and odds ratio, and Robins’ (3) IPW estimator for the odds ratio. We devote much of our attention to describing the implementation of the case-control-weighted TMLE for the risk difference.

Double robust methods are appealing to researchers because the investigators are afforded two opportunities to obtain an unbiased estimator of their parameter of interest. Consider, for example, a simple cross-sectional study with an outcome, binary exposure, and a vector of covariates. In order to construct the double robust estimator, we will first need an estimator for the conditional expectation of the outcome given exposure and covariates, as well as an estimator for the probability of the exposure given covariates. The precise definition of the double robust estimator will depend on whether it is derived on the basis of estimating functions (i.e., augmented IPW estimators) (17) or within the TMLE framework. Double robust estimators will generally be consistent if either the estimator of the conditional expectation of the outcome given exposure and covariates or the estimator of the probability of the exposure given covariates is consistently estimated. Additionally, double robust estimators will be efficient if both of these components are consistently estimated (4, 17).

Currently, a double robust augmented IPW estimator for case-control data does not yet exist. However, the TMLE framework allows one to both target a diverse array of parameters and analyze case-control data. These TMLEs are double robust efficient substitution estimators that can be applied to both observational and experimental data (4, 22). A detailed general discussion of TMLEs for various data structures compared with maximum likelihood substitution estimators of the g-formula (also referred to as G-computation estimators) (23), IPW estimators, and augmented IPW estimators is presented elsewhere (see chapter 6 in van der Laan and Rose (4)). A difference between the two double robust estimators, TMLEs and augmented IPW, is that augmented IPW estimators are not substitution estimators. Therefore, for example, an augmented IPW estimator can estimate a probability outside the bounds of the interval [0, 1]. A common example demonstrating the problem augmented IPW estimators face by not being substitution estimators is the exposure-specific mean, which is defined as the expectation of the outcome given exposure = 1 and covariates, averaged over the covariate strata. It is here, when the outcome is binary, that the augmented IPW estimator can yield values outside [0, 1]. Simulation studies comparing TMLE with other existing estimators have also been presented (4, 6, 24–30). In brief, the TMLE for the simple cross-sectional study with an outcome, binary exposure, and vector of covariates is constructed by first estimating the conditional expectation of the outcome given exposure and covariates. This estimate is updated using a covariate that reduces the bias of the initial estimate for the

target parameter. The covariate incorporates information from the probability of the exposure given covariates. Statistical theory shows us that the covariate moves the initial estimate in the direction of the target parameter, removing bias (4).

For unmatched case-control data, we use a case-control-weighted TMLE that requires knowledge of the marginal prevalence probability of the outcome from an external source (e.g., a cancer registry) to eliminate the bias induced by the sampling design. The prevalence probability within levels of measured exposures is not required. This prevalence probability can be either known or estimated from a previous cohort study, and the availability of such values is increasingly realistic given regional and national databases (5, 6). The prevalence probability is basic information about a population of interest in the sense that it is a proportion necessary to carefully define the population. When this type of information has not been readily available, epidemiologists have proposed and conducted a census of their target population to collect various population-level proportions prior to sampling—for example, in the Study of Physical Performance and Age-Related Changes in Sonomans, a longitudinal cohort study (31). The literature has also long been supportive of the use of the prevalence probability (32). Another existing case-control method uses the prevalence probability to eliminate the bias of the case-control sampling design by adding $\log(\text{prevalence probability}/(1 - \text{prevalence probability}))$ to the intercept of a correctly specified parametric logistic regression (33, 34). This regression function can then be used for prediction or as part of “plug-in” estimators of various effect parameters. A corresponding approach is used in randomized recruitment case-control designs (35).

By exploiting the prevalence probability in case-control-weighted TMLEs, we are also not limited to outcomes with a prevalence probability of less than 10%, or another “rare disease” cutoff, for our estimator to perform well, unlike the IPW methods. VanderWeele and Vansteelandt propose that the prevalence probability should be less than 10% for adequate performance of their estimator (21), and thus the application of the method is restricted to outcomes that do not exceed this threshold. The original IPW estimator for the odds ratio in case-control data published by Robins (3) requires a “rare” outcome (the example given is 3%). Note that none of the estimators discussed can remove prevalence-incidence bias—that is, the bias that is present if certain cases are not eligible for sampling because of disease severity.

For clarity of presentation, we focus on unmatched case-control studies using known prevalence probability. TMLEs for individually matched case-control studies have been presented (7), as has the use of estimated prevalence probabilities within a case-control-weighted TMLE (see page 39 in van der Laan (5)). Implementation of the TMLE for matched case-control data is very similar to that for unmatched case-control studies; the differences lie in the prevalence information required to adjust for the biased sampling (5, 7). There are also several double robust estimators available for effect estimation in so-called nested case-control studies, including estimating equation methods (36) and a TMLE (37). These designs are, in truth, a missing-data sampling problem versus a case-control sampling problem and are often referred to as 2-stage or 2-phase studies.

CASE-CONTROL-WEIGHTED TMLEs: ESTIMATORS OF MULTIPLE PARAMETERS

A benefit of the case-control-weighted TMLE framework is that we are able to target various parameters that were not previously available when analyzing case-control data. Investigators were largely limited to the conditional or marginal odds ratio, and now also the RERI approximated using odds ratios, which may not have been the parameter of greatest interest or the one that answered their research question. Ideally, the structure of the data (i.e., the biased case-control sampling study design) should not dictate the parameter one estimates. We present an example of use of the case-control-weighted TMLE for the simplest parameter, the marginal risk difference. The relative risk and odds ratio are only slightly more difficult to implement. The case-control-weighted TMLE has also previously been described for the relative risk and the odds ratio (5, 6). Let us define $X = (W, A, Y) \sim P_{X,0}$ as the unobserved full-data experimental unit X with outcome Y , binary exposure A , and vector of covariates W , and the true underlying distribution of interest $P_{X,0}$. (The subscript “ $X,0$ ” indicates the true full distribution versus a “possible” full distribution denoted by P_X .)

Uppercase letters represent random variables, and lowercase letters represent a specific value for that random variable. To be explicit, if all variables are discrete, which we use as an example in this paper, writing $P_{X,0}(W = w, A = a, Y = y)$ indicates that we are assigning a probability to the inputs (w, a, y) . We can succinctly define the risk difference (RD) by

$$\text{RD} = P_{W,0}[P_{X,0}(Y = 1|A = 1, W) - P_{X,0}(Y = 1|A = 0, W)], \quad (1)$$

where $P_{W,0}$ is the true probability distribution of the covariates W . An equivalent presentation of the parameter is

$$\text{RD} = \sum_w \left[\sum_y y P_{X,0}(Y = y|A = 1, W = w) - \sum_y y P_{X,0}(Y = y|A = 0, W = w) \right] P_{X,0}(W = w),$$

where

$$P_{X,0}(Y = y|A = a, W = w) = \frac{P_{X,0}(W = w, A = a, Y = y)}{\sum_y P_{X,0}(W = w, A = a, Y = y)}$$

is the conditional probability distribution of $Y = y$, given $A = a$, $W = w$, and

$$P_{X,0}(W = w) = \sum_{y,a} P_{X,0}(Y = y, A = a, W = w)$$

is the marginal probability distribution of $W = w$. Since we have case-control data, our outcome Y is binary. The prevalence probability is given as $P_{X,0}(Y = 1) = q_0$. Note that the prevalence probability is defined by the same probability distribution that defines the full data. Therefore, if we were

studying the effect of a toxin on cancer among US men in 2002, the prevalence probability would be defined by the prevalence in this population in the year 2002. For detailed discussions of the assumptions necessary for a causal interpretation of this parameter of interest (no unmeasured confounders, consistency, and no interference between subjects), we refer to previous material (4, 8, 9).

Suppose the observed data structure for an unmatched case-control study is described as sampling nC cases from the conditional distribution of (W, A) given $Y = 1$ and sampling nCo controls from the conditional distribution of (W, A) given $Y = 0$. The value $J = nCo/nC$ that we will use in the case-control weights is the average number of controls per case. For this data structure, the procedure for calculating the case-control-weighted TMLE is as follows.

Step 1. Assign the weight q_0 to each case and the weight $(1 - q_0)/J$ to each control. Each time we use weights in the case-control-weighted TMLE procedure, we will use these weights.

Step 2. Estimate the conditional distribution of the outcome $Y = 1$, given exposure and covariates, $P_{X,0}(Y = 1|A, W)$. One can use a case-control-weighted parametric logistic regression to estimate this function or any estimator that allows for weighted observations. For example, various machine learning procedures allow for observation weighting (4, 38–40). We focus on the case where an a priori parametric logistic regression is specified to estimate $P_{X,0}(Y = 1|A, W)$ using the weights defined in step 1. This estimate is denoted $P_{X,n}(Y = 1|A, W)$, where the subscript n indicates that it is an estimator. It is at this step that one could stop, plug this initial estimate of $P_{X,0}(Y = 1|A, W)$ into the parameter in equation 1, and then average over the case-control-weighted distribution of the covariates W . This would be the case-control-weighted G-computation estimator (5, 6), which does not enjoy the double robustness properties of the TMLE.

Step 3. Estimate the probability of the exposure given the covariates $P_{X,0}(A|W)$ using case-control-weighted logistic regression. (Other estimators can be used to estimate this function as well.) This estimate is denoted $P_{X,n}(A|W)$.

Step 4. Calculate the covariate denoted as

$$H(A, W) = \frac{I(A = 1)}{P_{X,n}(A = 1|W)} - \frac{I(A = 0)}{P_{X,n}(A = 0|W)}. \quad (2)$$

This covariate incorporates information from the estimate of the probability of the exposure given the covariates and will vary depending on the parameter estimated.

Step 5. Update the initial fit achieved in step 2 by holding the coefficients of $P_{X,n}(Y = 1|A, W)$ fixed, while estimating a coefficient ε for $H(A, W)$ using case-control-weighted maximum likelihood estimation in the following submodel:

$$P_{X,n}^{\text{updated}}(Y = 1|A, W) = \text{expit} \left(\log \left(\frac{P_{X,n}(Y = 1|A, W)}{1 - P_{X,n}(Y = 1|A, W)} \right) + \varepsilon H(A, W) \right).$$

Step 6. Estimate the parameter given in equation 1 by plugging in the updated estimate of $P_{X,n}(Y = 1|A, W)$ under $A = 1$

and $A = 0$ and averaging over the case-control-weighted distribution of the covariates W .

Standard errors can be calculated using the influence curve or bootstrapping methods (5, 6). Bootstrapping, where the case-control sample is resampled s times and the standard deviation of the bootstrap estimates of the parameter is used as an estimate of the standard error, can also be implemented. We refer readers to the next section and the Appendix for more information on using influence curves. Inference for TMLEs incorporating machine learning has also been discussed (4).

SIMULATION ILLUSTRATION

To illustrate the case-control-weighted TMLE methodology, we generated a simulated data set to represent the analysis of a single data set. Consider a study of liver cancer in China taking place in 2002, where prevalence of the disease was 157,932 per 1,285,500,000 persons or 1.23×10^{-4} , according to the International Agency for Research on Cancer (41). The main causes of liver cancer are chronic hepatitis B, chronic hepatitis C, aflatoxin exposure, and cirrhosis. Suppose we are interested in the effect of cirrhosis on liver cancer. We simulated a target population of 5 million subjects.

For the target population, male sex was simulated with a probability of 0.515, chronic hepatitis C was simulated with a probability of 0.038, chronic hepatitis B was simulated with a probability of 0.023, and living in a rural area versus an urban area was generated with a probability of 0.65. Age was distributed as a truncated normal distribution with a mean of 30 and a standard deviation of 10, truncated on the interval [18, 70]. Average alcohol consumption (in L/year) was generated from a truncated normal distribution with the following mean:

$$E_{X,0}(\text{alcohol}|\text{sex}) = 4 + 2 \times \text{sex},$$

with a standard deviation of 1.5 L/year and truncated on the interval [0, 20]. Sex was coded as 1 if male and 0 if female. The generation of aflatoxin exposure depended on both location (rural vs. urban) and age:

$$\begin{aligned} \text{logit}(P_{X,0}(\text{aflatoxin} = 1|\text{rural}, \text{age})) \\ = 0.9 - 1.9 \times \text{rural} + 0.985 \times \sqrt{\text{rural}} \\ + 0.0005 \times \text{age}. \end{aligned}$$

Binary smoking status (yes = 1, no = 0) was contingent on sex:

$$\text{logit}(P_{X,0}(\text{smoking} = 1|\text{sex})) = 3.05 + 3.407 \times \text{sex}.$$

The exposure cirrhosis depended on smoking status, chronic hepatitis B status (hepB), chronic hepatitis C status (hepC), age, alcohol consumption, and aflatoxin:

$$\begin{aligned} \text{logit}(P_{X,0}(\text{cirrhosis} = 1|\text{smoking}, \text{hepB}, \text{hepC}, \text{age}, \\ \text{alcohol}, \text{aflatoxin})) \\ = -4.4 + 2 \times \text{smoking} + 0.2 \times \text{hepB} + 0.1 \times \text{hepC} \\ + 0.002 \times \text{age} + 1.2 \times \text{alcohol} + 0.1 \times \text{aflatoxin}. \end{aligned}$$

Lastly, the binary outcome liver cancer had the following distribution, where it depended on sex, smoking status, chronic

hepatitis B status, chronic hepatitis C status, age, alcohol consumption, aflatoxin, and cirrhosis:

$$\begin{aligned} & \text{logit}(P_{X,0}(\text{liver cancer} = 1 | \text{cirrhosis, sex, smoking, hepB,} \\ & \quad \text{hepC, age, alcohol, aflatoxin})) \\ &= -9.8 + 0.4 \times \text{cirrhosis} + 0.05 \times \text{sex} + 0.002 \times \text{smoking} \\ & \quad + \text{hepB} + \text{hepC} + 0.02 \times \text{age} + 0.2 \times \text{alcohol} \\ & \quad + 0.01 \times \text{aflatoxin}. \end{aligned}$$

These values and the relationships between the variables were based on population data for China. For example, in 2002 it was estimated that 32% of the population of China smoked, with 60% of males smoking and only 4% of females (41–45). In our simulated target population, 32% of the population smoked, with 59% of males smoking and 5% of females smoking. The estimated prevalence probability in our simulated population was 1.45×10^{-4} , which was close to the population value of 1.23×10^{-4} given in the literature. However, we included some simplifications in constructing this simulated population for ease of presentation—for example, assuming that hepatitis B status was independent of sex.

We extracted a case-control sample from this target population, with 727 cases (diagnosed with liver cancer) and 2,908 controls, for an overall sample size of 3,050. We wished to estimate the parameter defined in equation 1, the risk difference, where Y = liver cancer, A = cirrhosis, and W = (sex, age, rural, hepC, hepB, alcohol, aflatoxin, smoking). The steps employed follow those described in the previous section.

Step 1. We assigned the weight $q_0 = 1.45 \times 10^{-4}$ to each case and the weight $(1 - q_0)/J = (1 - 1.45 \times 10^{-4})/4 = 2.50 \times 10^{-1}$, where we recall that J is the average number of controls per case.

Step 2. The conditional distribution of the outcome given exposure and covariates was estimated using a correctly specified parametric weighted logistic regression with the weights described in step 1. This fit was estimated by

$$\begin{aligned} & \text{logit}(P_n(\text{liver cancer} = 1 | \text{cirrhosis, sex, smoking, hepB,} \\ & \quad \text{hepC, age, alcohol, aflatoxin})) \\ &= -9.33 + 0.37 \times \text{cirrhosis} + 0.19 \times \text{sex} \\ & \quad - 0.04 \times \text{smoking} + 1.14 \times \text{hepB} + 1.07 \times \text{hepC} \\ & \quad + 0.02 \times \text{age} - 0.06 \times \text{alcohol} - 0.02 \times \text{aflatoxin}. \quad (3) \end{aligned}$$

Step 3. We estimated the probability of the exposure given the covariates with a mildly misspecified weighted logistic regression, again using the weights described in step 1. The variable sex, which was not a part of the true data-generating distribution, was included. This fit was given by

$$\begin{aligned} & \text{logit}(P_n(\text{cirrhosis} = 1 | \text{sex, smoking, hepB, hepC,} \\ & \quad \text{age, alcohol, aflatoxin})) \\ &= -4.50 + 0.16 \times \text{sex} + 1.78 \times \text{smoking} \\ & \quad - 0.14 \times \text{hepB} + 0.46 \times \text{hepC} + 0.02 \times \text{age} \\ & \quad + 0.13 \times \text{alcohol} + 0.04 \times \text{aflatoxin}. \end{aligned}$$

Step 4. The covariate was calculated for each subject using equation 2 and the estimates generated by the fit given in step 3.

Step 5. We updated the initial fit from step 2 by holding the coefficients of equation 3 fixed while estimating a coefficient ε for $H(A, W)$ using case-control-weighted maximum likelihood estimation.

Step 6. The parameter given in equation 1 was estimated by plugging the updated estimate from step 5 into the formula and averaging over the case-control-weighted distribution of the covariates W . The estimated value was 6.34×10^{-5} , and the true value of the risk difference was 5.70×10^{-5} . An estimate of the standard error for the risk difference was calculated using estimated influence curves (see Appendix). The estimated standard error was 3.27×10^{-5} , with $P = 0.053$ and a 95% confidence interval of $[-7.57 \times 10^{-7}, 1.28 \times 10^{-4}]$.

We focused on the risk difference in this presentation, since this paper is largely a tutorial, and it is the simplest parameter to implement and understand. Additionally, it highlights the fact that TMLE allows for the estimation of a variety of parameters. We recognize that other parameters may be more clinically useful and remind the reader that the investigators may have decided a priori that they were interested in another parameter, including the relative risk or the odds ratio, which we can also estimate.

SUMMARY

Additional methods beyond those discussed by VanderWeele and Vansteelandt (21) are available for the estimation of marginal parameters in case-control data. These estimators include case-control-weighted TMLEs (5, 6) for the risk difference, relative risk, and odds ratio and the first IPW estimator for case-control data, published by Robins (3), which estimates the odds ratio. VanderWeele and Vansteelandt presented an IPW estimator of the RERI parameter for case-control studies (21). Their proposed method does not address other parameters. Researchers wishing to conduct and analyze case-control studies will benefit from understanding all of the currently available tools.

In this commentary, we have described the estimation procedure for the risk difference using a case-control-weighted TMLE in unmatched data. A case-control-weighted TMLE for the RERI has not yet been presented in the literature. Such an estimator could be constructed by targeting the multidimensional parameter for the two exposures, but we leave this derivation for future work. We also note that if the RERI is truly the parameter of interest, a TMLE estimate can be derived that uses estimates of the relative risks and does not need to be approximated using odds ratios as done by VanderWeele and Vansteelandt. By requiring knowledge of the prevalence probability in case-control-weighted TMLEs, or an estimate of this probability, there is also no cutoff value for the outcome prevalence to ensure a high level of performance, unlike the currently proposed IPW estimators. Ascertaining the prevalence probability should be a high priority when possible, if it is not already available through a previous study, census, or regional database. In summary, case-control-weighted TMLEs are double robust efficient substitution estimators and thus have desirable asymptotic and finite sample performance properties. Researchers may wish to consider their use in future studies, along with the IPW methods of Robins (3) and VanderWeele and Vansteelandt (21).

ACKNOWLEDGMENTS

Author affiliations: Department of Health Care Policy, Harvard Medical School, Boston, Massachusetts (Sherri Rose); and Division of Biostatistics, School of Public Health, University of California, Berkeley, Berkeley, California (Mark van der Laan).

This work was supported by the National Science Foundation (grant DMS-1103901 to S.R.) and the National Institutes of Health (grant R01 A1074345-01 to M.v.d.L.).

Conflict of interest: none declared.

REFERENCES

1. Breslow NE, Day NE. *Statistical Methods in Cancer Research. Vol. 1. The Analysis of Case-Control Studies*. (IARC Scientific Publication no. 32). Lyon, France: International Agency for Research on Cancer; 1980.
2. Rothman K, Greenland S. *Modern Epidemiology*. 2nd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 1998.
3. Robins JM. Choice as an alternative to control in observational studies: comment. *Stat Sci*. 1999;14(3):281–293.
4. van der Laan MJ, Rose S. *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York, NY: Springer Publishing Company; 2003.
5. van der Laan MJ. Estimation based on case-control designs with known prevalence probability. *Int J Biostat*. 2008;4(1): Article 17.
6. Rose S, van der Laan MJ. Simple optimal weighting of cases and controls in case-control studies. *Int J Biostat*. 2008;4(1): Article 19.
7. Rose S, van der Laan MJ. Why match? Investigating matched case-control study designs with causal effect estimation. *Int J Biostat*. 2009;5(1):Article 1.
8. Robins JM. Marginal structural models versus structural nested models as tools for causal inference. In: Halloran ME, Berry D, eds. *Statistical Models in Epidemiology, the Environment, and Clinical Trials*. New York, NY: Springer-Verlag New York; 1999:95–134.
9. Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*. 2000;11(5): 561–570.
10. Hernán MA, Brumback BA, Robins JM. Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for repeated measures. *Stat Med*. 2002;21(12): 1689–1709.
11. Cole SR, Hernán MA, Robins JM, et al. Effect of highly active antiretroviral therapy on time to acquired immunodeficiency syndrome or death using marginal structural models. *Am J Epidemiol*. 2003;158(7):687–694.
12. Bodnar LM, Davidian M, Siega-Riz AM, et al. Marginal structural models for analyzing causal effects of time-dependent treatments: an application in perinatal epidemiology. *Am J Epidemiol*. 2004;159(10):926–934.
13. Tager IB, Haight T, Sternfeld B, et al. Effects of physical activity and body composition on functional limitation in the elderly: application of the marginal structural model. *Epidemiology*. 2004;15(4):479–493.
14. Haight T, Tager I, Sternfeld B, et al. Effects of body composition and leisure-time physical activity on transitions in physical functioning in the elderly. *Am J Epidemiol*. 2005; 162(7):607–617.
15. Cole SR, Hernán MA, Anastos K, et al. Determining the effect of highly active antiretroviral therapy on changes in human immunodeficiency virus type 1 RNA viral load using a marginal structural left-censored mean model. *Am J Epidemiol*. 2007; 166(2):219–227.
16. Petersen ML, Wang Y, van der Laan MJ, et al. Pillbox organizers are associated with improved adherence to HIV antiretroviral therapy and viral suppression: a marginal structural model analysis. *Clin Infect Dis*. 2007;45(7): 908–915.
17. van der Laan MJ, Robins JM. *Unified Methods for Censored Longitudinal Data and Causality*. New York, NY: Springer Publishing Company; 2003.
18. Neugebauer R, van der Laan M. Why prefer double robust estimators in causal inference? *J Stat Plan Inference*. 2005; 129(1-2):405–426.
19. Ertefaie A, Stephens DA. Comparing approaches to causal inference for longitudinal data: inverse probability weighting versus propensity scores. *Int J Biostat*. 2010;6(2):14.
20. Petersen ML, Porter KE, Gruber S, et al. Diagnosing and responding to violations in the positivity assumption. *Stat Methods Med Res*. 2012;21(1):31–54.
21. VanderWeele TJ, Vansteelandt S. A weighting approach to causal effects and additive interaction in case-control studies: marginal structural linear odds models. *Am J Epidemiol*. 2011; 174(10):1197–1203.
22. van der Laan MJ, Rubin D. Targeted maximum likelihood learning. *Int J Biostat*. 2006;2(1):Article 11.
23. Robins JM. A new approach to causal inference in mortality studies with sustained exposure periods—application to control of the healthy worker survivor effect. *Math Model*. 1986; 7(9–12):1393–1512.
24. Gruber S, van der Laan MJ. An application of collaborative targeted maximum likelihood estimation in causal inference and genomics. *Int J Biostat*. 2010;6(1):Article 18.
25. Stitelman OM, van der Laan MJ. Collaborative targeted maximum likelihood for time-to-event data. *Int J Biostat*. 2010; 6(1):Article 21.
26. Gruber S, van der Laan MJ. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. *Int J Biostat*. 2010;6(1):Article 26.
27. Wang H, Rose S, van der Laan MJ. Finding quantitative trait loci genes with collaborative targeted maximum likelihood learning. *Stat Probab Lett*. 2011;81(7):792–796.
28. Porter KE, Gruber S, van der Laan MJ, et al. The relative performance of targeted maximum likelihood estimators. *Int J Biostat*. 2011;7(1):Article 31.
29. Muñoz ID, van der Laan MJ. Population intervention causal effects based on stochastic interventions. *Biometrics*. 2011; 68(2):541–549.
30. Jamshidian F, Hubbard AE, Jewell NP. Accounting for perception, placebo and unmasking effects in estimating treatment effects in randomized clinical trials [published online ahead of print September 8, 2011]. *Stat Methods Med Res*. (doi:10.1177/0962280211413449).
31. Tager IB, Hollenberg M, Satariano WA. Association between self-reported leisure-time physical activity and measures of cardiorespiratory fitness in an elderly population. *Am J Epidemiol*. 1998;147(10):921–931.
32. Cornfield J. A method of estimating comparative rates from clinical data: applications to cancer of the lung, breast, and cervix. *J Natl Cancer Inst*. 1951;11(6):1269–1275.
33. Anderson JA. Separate sample logistic discrimination. *Biometrika*. 1972;59(1):19–35.
34. Prentice RL, Breslow NE. Retrospective studies and failure time models. *Biometrika*. 1978;65(1):153–158.

35. Weinberg CR, Sandler DP. Randomized recruitment in case-control studies. *Am J Epidemiol*. 1991;134(4):421–432.
36. Wang W, Scharfstein D, Tan Z, et al. Causal inference in outcome-dependent two-phase sampling designs. *J R Stat Soc Series B Stat Methodol*. 2009;71(5):947–969.
37. Rose S, van der Laan MJ. A targeted maximum likelihood estimator for two-stage designs. *Int J Biostat*. 2011;7(1): Article 17.
38. Venables WN, Ripley BD. *Modern Applied Statistics with S*. 4th ed. New York, NY: Springer-Verlag New York; 2002.
39. van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol*. 2007;6(1):Article 25.
40. Rose S. Mortality risk score prediction in an elderly population using machine learning. *Am J Epidemiol*. 2013;177(5): 443–452.
41. Ferlay J, Shin HR, Bray F, et al, eds. *GLOBOCAN 2008: Cancer Incidence and Mortality Worldwide. IARC CancerBase No. 10*. Lyon, France: International Agency for Research on Cancer; 2010.
42. Qian GS, Ross RK, Yu MC, et al. A follow-up study of urinary markers of aflatoxin exposure and liver cancer risk in Shanghai, People's Republic of China. *Cancer Epidemiol Biomarkers Prev*. 1994;3(1):3–10.
43. Liu J, Fan D. Hepatitis B in China. *Lancet*. 2007;369(9573): 1582–1583.
44. Pfizer Inc. *The Burden of Cancer in Asia*. New York, NY: Pfizer Inc.; 2008. (http://www.pfizer.com/files/products/cancer_in_asia.pdf). (Accessed April 1, 2012).
45. Liu F, Chen K, He Z, et al. Hepatitis C seroprevalence and associated risk factors, Anyang, China. *Emerg Infect Dis*. 2009; 15(11):1819–1822.

APPENDIX

The estimated influence curve is an alternative method to the bootstrap for calculating standard errors. For detailed

theory about influence curves, we refer the reader to other literature (4, 6, 7, 22). The intuition behind the influence curve is that it gives us an indication of what happens to our estimator when 1 value is changed. The influence curve (IC) for the risk difference (RD) in a study without biased sampling is given by

$$\begin{aligned} \text{IC}_n(W_i, A_i, Y_i) = & \left(\frac{I(A_i = 1)}{P_n(A = 1|W_i)} - \frac{I(A_i = 0)}{P_n(A = 0|W_i)} \right) \\ & \times (Y - E_n(Y|A_i, W_i)) + E_n(Y|A = 1, W_i) \\ & - E_n(Y|A = 0, W_i) - \text{RD}_n, \end{aligned}$$

where $E_n(Y|A_i, W_i)$ is the updated fit from step 5. With the addition of the biased sampling, the influence curve for the risk difference, as described in this paper, is given by

$$\begin{aligned} \text{IC}_{n,q0}(W_i, A_i, Y_i) = & q_0 \text{IC}_n(W_i, A_i, 1) \\ & + (1 - q_0) \frac{1}{J} \sum_{j=1}^J \text{IC}_n(W_i, A_i, 0). \end{aligned}$$

Once this influence curve is calculated, we use it to estimate the sample variance of the estimated influence curve values:

$$S^2(\text{IC}_n) = \frac{1}{n} \sum_{i=1}^n (\text{IC}_n(o_i) - (\overline{\text{IC}_n}))^2.$$

We can then calculate the standard error by taking the square root of the estimated sample variance divided by sample size. Confidence intervals and *P* values can be constructed using this estimate of the standard error.