Epidemiologic studies often aim to estimate the odds ratio for the association between a binary exposure and a binary disease outcome. Because confounding bias is of serious concern in observational studies, investigators typically estimate the adjusted odds ratio in a multivariate logistic regression which conditions on a large number of potential confounders. It is well known that modeling error in specification of the confounders can lead to substantial bias in the adjusted odds ratio for exposure. As a remedy, Tchetgen Tchetgen et al. (Biometrika. 2010;97(1):171–180) recently developed so-called doubly robust estimators of an adjusted odds ratio by carefully combining standard logistic regression with reverse regression analysis, in which exposure is the dependent variable and both the outcome and the confounders are the independent variables. Double robustness implies that only one of the 2 modeling strategies needs to be correct in order to make valid inferences about the odds ratio parameter. In this paper, I aim to introduce this recent methodology into the epidemiologic literature by presenting a simple closed-form doubly robust estimator of the adjusted odds ratio for a binary exposure. A SAS macro (SAS Institute Inc., Cary, North Carolina) is given in an online appendix to facilitate use of the approach in routine epidemiologic practice, and a simulated data example is also provided for the purpose of illustration.

Abbreviation: SE, standard error.

Many epidemiologic studies aim to estimate, using observational data, an adjusted odds ratio for a binary exposure as it relates to the risk of a binary disease outcome, conditional on a moderate-to-large number of potential confounders. Logistic regression is widely used as the standard analytical tool for estimating the exposure-disease adjusted odds ratio, which we denote \( \exp(\psi^*) \), by fitting, using maximum likelihood estimation, a working model

\[
\logit\{ \Pr(Y = 1|A, L; \psi, \eta) \} = \eta_1 + \eta_2^L + \psi A \quad \text{(Model 1)}
\]

for the conditional probability of the occurrence of the outcome given exposure level \( A \) and confounders \( L \), where \( \eta = (\eta_1, \eta_2^L)' \). Throughout, we assume that, as encoded in model 1, the effect of exposure is homogeneous on the odds ratio scale; that is, we assume that for any individual with covariates \( L \), the log odds ratio for the exposure is equal to the constant \( \psi^* \), the same for all values of \( L \), so that \( \psi^* \) is the true value of \( \psi \) in model 1. Note that if unmeasured confounding is present, \( \psi^* \) will generally fail to have a causal interpretation. However, it remains a well-defined summary measure under model 1 of the partial association between exposure and outcome after adjustment for the measured covariates \( L \), and it often constitutes the primary target of inference in epidemiologic analyses. Our discussion concerns inference about the parameter \( \psi^* \) regardless of whether or not it has a meaningful causal interpretation, that is, whether or not \( L \) is sufficiently rich to fully adjust for confounding.
Unfortunately, the estimator of Tchetgen Tchetgen et al. (3) and a simple closed-form doubly robust estimator is known to hold. Let 

\[ \alpha = (\alpha_1, \alpha_2)^T \]. Crucially, the parameter \( \psi \) is shared between models 1 and 2, reflecting the following key property of odds ratios:

\[
\exp(\psi) = \frac{\Pr(Y = 1|A = 1, L)\Pr(Y = 0|A = 0, L)}{\Pr(Y = 1|A = 0, L)\Pr(Y = 0|A = 1, L)}
\]

\[
= \frac{\Pr(A = 1|Y = 1, L)\Pr(A = 0|Y = 0, L)}{\Pr(A = 1|Y = 0, L)\Pr(A = 0|Y = 1, L)} \quad \text{(Model 3)}
\]

Therefore, models 1 and 2 represent two genuine (i.e., independent) opportunities to correctly estimate the exposure effect, in the sense that the 2 working models are perfectly compatible for any value of \( \alpha \) and \( \eta \). However, even if model 3 is known to hold, in practice it is generally impossible to know with certainty which, if either, of regression models 1 and 2 is correctly specified. In fact, one may wish to combine these two distinct strategies for estimation into a single overall strategy that is guaranteed in large samples, to deliver a correct estimate of \( \psi^* \) provided that one of the 2 models (1 and 2) is correctly specified, without necessarily knowing which is correct. Tchetgen Tchetgen et al. (3) recently developed a large class of estimators with precisely this desirable property. Stated more precisely, the methods of Tchetgen Tchetgen et al. (3) yield a specific class of estimators of \( \psi^* \) that are asymptotically unbiased if one, but not necessarily both, of the following is true:

1. The working model \( \Pr(A = 1|Y = 0, L; \alpha) \) is correctly specified even if \( \Pr(Y = 1|A = 0, L; \eta) \) is incorrectly specified.
2. The working model \( \Pr(Y = 1|A = 0, L; \eta) \) is correctly specified even if the working model \( \Pr(A = 1|Y = 0, L; \alpha) \) is incorrectly specified.

Unfortunately, the estimator of Tchetgen Tchetgen et al. (3) is not easily obtained without special software, thus seriously impeding its routine use. This computational challenge inspired the SAS macro (SAS Institute Inc., Cary, North Carolina) of Tchetgen Tchetgen and Rotnitzky (4), which implements the optimal doubly robust estimator via an iterative procedure which they called the ProRetroSpec algorithm. Here, the above computational challenge is further addressed in the simple case of a binary exposure, and a simple closed-form doubly robust estimator is provided which is easy to compute using standard software.

Suppose that independent and identically distributed data on the variables \( (A, Y, L) \) are observed in \( n \) individuals for whom the homogeneous odds ratio model, model 3, is known to hold. Let \( \hat{\alpha} = \Pr(A = 1|Y = 0, L; \hat{\alpha}) \), where \( \hat{\alpha} \) denotes the maximum likelihood estimator of \( \alpha \) using data

\[
\hat{\psi}(w) = \log \frac{\sum W_i A_i Y_i (1 - \hat{P}_i)(1 - \hat{B}_i)}{\sum W_i (1 - A_i)Y_i (1 - \hat{P}_i)(1 - \hat{B}_i) - (1 - A_i)(1 - Y_i)(1 - \hat{P}_i)(1 - B_i)}
\]

For a fixed choice (possibly estimated) of \( w(\cdot) \), the resulting estimator \( \hat{\psi}(w) \) is doubly robust and thus converges to \( \psi^* \) (in probability) with increasing sample size, provided that either condition 1 or condition 2 holds, but not necessarily both. A formal argument establishing double robustness of \( \hat{\psi}(w) \) is given in the Web Appendix (available at http://aje.oxfordjournals.org/) for completeness. The choice of the weight function \( w(\cdot) \) affects only efficiency, and the optimal choice of the weight \( w(\cdot) \) is easily inferred from a result presented in the paper by Tchetgen Tchetgen et al. (3); it is beyond the scope of this article. Sample SAS code for the simple estimator \( \psi = \hat{\psi}(1) \) is provided in the Web Appendix, along with code for computing standard errors that correctly account for the variation in \( \hat{P}_i \) and \( \hat{B}_i \).

**A DATA ILLUSTRATION**

Here I briefly illustrate the application of the doubly robust macro in a simulated data set. For this purpose, we generate independent and identically distributed data on 2,000 individuals under models 1 and 2 with covariates \( L = (1, L_1, L_2, L_3 = L_1 \times L_2) \), where \( (L_1, L_2) \) are independent standard normal, \( \psi^* = 0.5 \), and regression coefficients \( \alpha = (0.5, -0.3, 0.5, -0.5)^T \), \( \eta = (0.5, 0.5, -0.5, -0.5) \).

To illustrate the presence of bias under model misspecification of model 1, we obtain 2 estimates of \( \psi^* \) using the standard (prospective) logistic regression model, model 1, one which includes the interaction \( L_3 \) in \( L \) and another which does not. Similarly, we obtain 2 estimates of \( \psi^* \) in the reverse regression model, model 2, one which includes the interaction \( L_3 \) in \( L \) and another which does not. Table 1 summarizes results for these estimators.

Next, recall that the doubly robust estimator uses both model 1 and model 2 to estimate \( \psi^* \), and we aim to show that only one model needs to be correct for valid inference. Thus, we obtained 3 doubly robust estimators. The first

<table>
<thead>
<tr>
<th>Interaction</th>
<th>No interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>OR SE</td>
<td>OR SE</td>
</tr>
<tr>
<td>0.56 0.11</td>
<td>0.76 0.10</td>
</tr>
<tr>
<td>0.53 0.11</td>
<td>0.75 0.10</td>
</tr>
</tbody>
</table>

**Table 1.** Prospective and Reverse Logistic Regression Estimates of \( \psi \)

Abbreviations: OR, odds ratio; SE, standard error.
estimator, $\hat{\psi}_{\text{pr}}^{\text{correct}} = 0.54$ (standard error (SE), 0.11), used correct models for models 1 and 2. The second estimator, $\hat{\psi}_{\text{dr}}^{\text{correct}} = 0.55$ (SE, 0.10), used the correct outcome regression model (model 1) but misspecified the exposure model (model 2) by leaving out the interaction, while $\hat{\psi}_{\text{dr}}^{\text{correct}} = 0.56$ (SE, 0.10) was obtained under the reversed situation, that is, by misspecifying model 1 but correctly specifying model 2.

As one would expect, both logistic models, models 1 and 2, are consistent under correct model specification and are severely biased when the interaction between confounders is omitted. This example nicely illustrates the doubly robust property, since the doubly robust estimator is consistent in the absence of modeling error, and it agrees with the outcome regression estimate when the latter is consistent; it similarly agrees with the reverse regression when the latter is consistent. Therefore, this example provides empirical confirmation of the theoretical property that the doubly robust approach is consistent as long as at least one of models 1 and 2 is correct, without knowing which one holds. As expected, when both models 1 and 2 are incorrectly specified, the doubly robust estimator produces a biased estimate, $\hat{\psi}_{\text{dr}}^{\text{wrong}} = 0.77$ (SE, 0.11).

For completeness, we provide an example of the macro call used to estimate $\hat{\psi}_{\text{dr}}^{\text{correct}}$:

```%
\text{dr_odds(data=}_\text{cero_}, \text{y=}_\text{outcome_},
\text{a=}_\text{exposure_}, \text{oddsY=}_11\text{ }_12_\text{,}
\text{oddsA=}_11\text{ }_12\text{ }_13_\text{, sample=}_2000),
```

where the first argument to the macro gives the name of the data set which contained the variables (outcome, exposure, 11, 12, 13); the second argument specifies the outcome variable; the third argument specifies the exposure variable; the fourth argument lists the covariates to be included in the outcome regression model, model 1; the fifth argument lists the covariates to be included in the exposure model, model 2; and the final argument sets the sample size.

**FINAL REMARKS**

In closing, note that the doubly robust methodology described herein generalizes to polytomous and continuous exposures and can also incorporate effect modification of the exposure-outcome odds ratio by components of $L$, although closed-form estimators as obtained herein are generally not available in such more general settings (3). Additionally, because the odds ratio effect measure generally remains identified under a (matched) case-control design, the doubly robust methodology described herein equally applies under such outcome-dependent sampling designs (4).

**ACKNOWLEDGMENTS**

Author affiliation: Departments of Epidemiology and Biostatistics, Harvard University, Boston, Massachusetts (Eric J. Tchetgen Tchetgen).

This work was funded by National Institutes of Health grant R21ES019712-01.

Conflict of interest: none declared.

**REFERENCES**