Targeted Estimation of Marginal Absolute and Relative Associations in Case-Control Data: An Application in Social Epidemiology

M. Pearl
Laura Balzer
J. Ahern

Available at: https://works.bepress.com/laura_balzer/34/
Targeted Estimation of Marginal Absolute and Relative Associations in Case–Control Data

An Application in Social Epidemiology

Michelle Pearl, Laura Balzer, and Jennifer Ahern

Background: Case–control studies are useful for rare outcomes, but typical analyses limit investigators to parametric estimation of conditional odds ratios. Several methods exist for obtaining marginal risk differences and risk ratios in a case–control setting, including a recently described semiparametric targeted approach optimized for rare outcomes.

Methods: Using case–control data from a study of neighborhood poverty and very preterm birth, we demonstrate estimation of marginal risk differences and risk ratios and compare a parametric substitution estimator based on maximum likelihood estimation with targeted maximum likelihood estimation (TMLE), and a refinement of TMLE for rare outcomes that incorporates bounds on the conditional risk.

Results: In this illustration, living in a neighborhood with high poverty was associated with a higher risk of very preterm birth for white women. The estimated risk differences (cases/100) were 0.6 (95% confidence interval [CI]: 0.1, 1.1) from maximum likelihood estimation, 0.5 (95% CI: −1.1, 2.1) from TMLE, and 0.5 (95% CI: 0.0, 1.0) from the rare outcomes refinement. The rare outcomes refinement, which incorporates knowledge that the conditional risk is small, produced more precise estimates than TMLE. A similar pattern was observed for the relative risk.

Conclusion: Absolute and relative associations estimated from case–control data using a semiparametric targeted approach allow the scientist to determine the analysis and avoid unwarranted parametric assumptions. A rare outcomes refinement provided more precise estimates than TMLE, and thus is well suited for the study of rare outcomes.

Epidemiology 2016;27: 512–517
data on the underlying cohort is unavailable. Finally, cross validation can be used to select the upper bound if estimates are unavailable, or to select among a range of possible upper bounds indicated by other sources.22

We illustrate estimation of the marginal RD and risk ratio (RR) applying a weighting scheme developed by van der Laan12 to account for case–control sampling. We compare a parametric substitution estimator based on MLE with TMLE,17,19 and a rare outcomes targeted minimum loss-based estimator.21,22 To our knowledge, this is the first application of the rare outcomes refinement to case–control data.

METHODS

To illustrate implementation of MLE, TMLE, and the rare outcomes refinement in a case–control setting, we use data from three race/ethnicity specific cumulative case–control samples drawn from a cohort of 178,296 singleton live births. The goal was to assess the marginal RD and RR for the relationship between living in a poor neighborhood and risk of very preterm birth. The case–control samples were originally developed for the purpose of genetic analysis of stored specimens, described elsewhere.23 A total of 428 non-Hispanic white, 440 Mexican-American, and 247 Black very preterm birth cases (<32 weeks gestation) were identified based on gestational age from prenatal screening records linked to live birth records.23 Controls were term deliveries of normal birth weight infants, excluding complications of labor and delivery and birth defects.23 Addresses from birth records were geocoded and linked to 2000 Census data at the tract level. The study was approved by the California Committee for the Protection of Human Subjects. An index of neighborhood poverty was created based on factor analysis of percent below poverty, median household income, and percent on public assistance from all census tracts within the study area. For each racial/ethnic group, tracts above the median poverty level among controls in that group were considered high poverty. Confounders from birth records included maternal age, education, parity, and public insurance coverage for prenatal care or delivery. Three additional neighborhood-level indices were included as confounders in the analysis of the relation between neighborhood poverty and very preterm birth. Analyses were stratified by race/ethnicity to reflect the stratified sampling design, and to examine racial/ethnic differences commonly found in relations of neighborhood characteristics with birth outcomes.24–29 Analyses were conducted in R version 3.1.0 (The R Foundation for Statistical Computing, Vienna, Austria, 2014).30

The marginal risk of very preterm birth under the exposure (residence in a high-poverty neighborhood) and no exposure (residence in a low-poverty neighborhood), after adjusting for measured confounders, is represented within each racial/ethnic group (superscript r) as

\[
\text{Risk(exposed)}^r = E[P(Y = 1| A = 1,W^r)] \quad \text{and} \\
\text{Risk(unexposed)}^r = E[P(Y = 1| A = 0,W^r)]
\]

where \( Y \) is an indicator of very preterm birth, \( A \) is an indicator of residence in a high-poverty neighborhood (1 = above median poverty index, 0 = below median), and \( W^r \) is the vector of measured confounders.13,31 The expectation (\( E \)) averages over the population distribution of the confounders in each racial/ethnic group and thereby standardizes the conditional risk with respect to the group characteristics. On the additive and relative scales, these risks can be compared:

\[
\text{Risk difference}^r = E[P(Y = 1| A = 1,W^r)] - E[P(Y = 1| A = 0,W^r)] \\
\text{Risk ratio}^r = E[P(Y = 1| A = 1,W^r)]/E[P(Y = 1| A = 0,W^r)]
\]

These measures are the differences and ratios of the race/ethnicity specific risk of very preterm birth under the two exposure conditions, averaged with respect to the population confounder distribution of each racial/ethnic group.

For a cohort study design, there are several well-established methods to estimate the above parameters. Examples include inverse probability of treatment weighted estimators, augmented-inverse probability of treatment weighted estimators, and substitution estimators based on MLE and TMLE.17,32–34 Adapting these estimators for use with case–control samples can be accomplished by weighting each case by the population prevalence of the outcome and by weighting each control by (1-prevalence) divided by the ratio of the number of controls to the number of cases.12,15,19

Estimates of the population prevalence of very preterm birth for each race/ethnicity were obtained from the underlying cohort. Details of the sampling scheme and the weights are presented in the Supplemental Digital Content (eTable 1; http://links.lww.com/ED/B34). After applying this weighting scheme and excluding women with missing confounder data (5% excluded), we estimated the associations of interest for each racial/ethnic group with parametric MLE, semiparametric TMLE, and rare outcomes TMLE.21

To implement case–control weighted MLE, we first used weighted logistic regression to estimate the conditional probability of very preterm birth, given the exposure and confounders \( P(Y = 1| A = 1,W^r) \). Our a-priori-specified regression model included main terms for the exposure and confounders. Based on the regression coefficient estimates, we predicted each mother’s probability of very preterm birth, setting the exposure to be living in a poor neighborhood, given her observed confounders \( P_r(Y = 1|A = 1,W^r) \), where subscript \( n \) denotes an estimate. This was repeated, setting the exposure to be not living in a poor neighborhood, \( P_r(Y = 1|A = 0,W^r) \). Averaging these predicted probabilities over the weighted distribution of confounders in each racial/ethnic group yielded estimates of the marginal risk of very preterm birth under the exposure and the marginal risk under no exposure, respectively. The difference in these estimates provided an estimate on the absolute scale (marginal RD) of the association.
between living in a high-poverty neighborhood and the risk of very preterm birth. Likewise, the ratio of these estimates provided an estimate of the relative scale (marginal RR) of the association of living in a high-poverty neighborhood and the risk of very preterm birth. Confidence intervals were obtained from nonparametric bootstrap with 1000 iterations, sampling conditionally on case–control status.  

We next applied TMLE, which combines semiparametric estimation with an additional targeting step to reduce bias and obtain reliable inference. The targeting step involves estimating the exposure mechanism, which is the conditional probability of living in a high-poverty neighborhood, given the covariates. TMLE is double robust: if either the outcome regression or the exposure mechanism is consistently estimated, the estimated measures of association will be consistent (i.e., converge to the truth). In contrast, parametric MLE relies completely on specifying the correct functional form for the outcome regression. The TMLE is also locally efficient: if both the outcome regression and the exposure mechanism are correctly specified, the estimator will achieve the lowest possible variance among a large class of semiparametric estimators. The cost for TMLE’s semiparametric approach can be some loss of precision (relative to a correctly specified parametric regression for the outcome regression).  

To implement TMLE, we first estimated the conditional risk of very preterm birth, given the exposure and covariates \( P(Y = 1|A,W) \) with SuperLearner. SuperLearner is a data-adaptive method that uses cross-validation (i.e., sample splitting) to build the best combination of estimates from candidate algorithms. Our library of candidate algorithms was selected for interpretability and ability to incorporate analytic weights. The library included weighted main terms logistic regression (identical to case–control weighted MLE), weighted logistic regression with all possible pairwise interactions, stepwise weighted logistic regression, and the simple weighted mean. For the targeting step, the exposure mechanism (i.e., propensity score) \( P(A = 1|W) \) was estimated with SuperLearner, using the same library of weighted algorithms. Information in the estimated exposure mechanism was then used to update our initial estimates of the conditional risk of very preterm birth \( P_n(Y = 1|A,W) \) (see eAppendix 1, Supplemental Digital Content 1, which includes step-by-step implementation and a link to the \( R \) code; http://links.lww.com/ED/B34). Using the targeted estimates, we obtained each mother’s predicted probability of very preterm birth, setting the exposure to be living in a poor neighborhood and given her observed confounders; and her predicted probability of very preterm birth, setting the exposure to be not living in a poor neighborhood and given her observed confounders. By averaging these predicted outcomes, we obtained estimates of the marginal risk under the exposure, the marginal risk under no exposure, the marginal RD, and the marginal RR. As above, confidence intervals were based on the bootstrap with 1000 iterations and sampling conditionally on disease status.

For rare outcomes, Balzer et al recently introduced a refinement of the TMLE algorithm. The rare outcomes targeted minimum loss-based estimator is analogous to the standard TMLE procedure, but incorporates known or estimated bounds on the conditional probability of the outcome, given the exposure and covariates \( P(Y=1|A,W) \). These bounds can improve stability and power in finite samples. For case–control data, in particular, the rare outcomes refinement can help recover some of the precision lost by using a semiparametric approach. We selected a lower bound of 0 and an upper bound of 4 times the race/ethnic marginal prevalence. These bounds were selected based on the range of very preterm birth prevalence within covariate subgroups for each racial/ethnic group in the underlying cohort. The binary outcome (very preterm birth) was then transformed by subtracting off the lower bound (0 in this case) and dividing by the difference between the upper and lower bound. This transformed outcome was used for initial estimation with SuperLearner, targeting, and inference based on the bootstrap. The final estimates were transformed back to the original scale. We refer the reader to Balzer et al. for step-by-step implementation and accompanying R code.

**RESULTS**

The table presents results from the three estimation methods. For white women, the MLE-based estimates of the RD (RD [cases/100] = 0.6, 95% CI [0.1, 1.1]) and risk ratio (RR = 2.4 [1.2, 4.6]) suggested there was an association between living in a high-poverty neighborhood and the risk of very preterm birth. The estimates from TMLE (RD = 0.5 [−1.1, 2.1], RR = 2.0 [0.8, 5.1]) were in the same direction, but more modest and with confidence intervals that were about twice as wide. In general, differences in results between these methods are due in part to using a parametric main terms regression model in MLE versus data-adaptive methods in TMLE, as well as the targeting step in TMLE, which harnesses information in the exposure mechanism to reduce bias for the parameter of interest. The bootstrap confidence intervals for TMLE were substantially wider than those of MLE, as expected due to TMLE’s semiparametric approach; specifically, the estimated variance of the MLE for the RD was 7.4 × 10^-6, while the estimated variance of the TMLE was 6.6 × 10^-5. By incorporating additional knowledge about the rareness of the conditional risk of the outcome, much of the precision was recovered. The estimated variance of the rare outcomes refinement for the RD was 6.0 × 10^-6 and nearly order of magnitude smaller than the standard TMLE algorithm, while the point estimates were similar to TMLE (RD = 0.5 [0.0, 1.0], RR = 2.0 [1.1, 3.6]).

Among Hispanic women, the three estimation methods produced similar, imprecise estimates. Again, TMLE resulted in the widest confidence intervals, and the rare outcomes refinement recovered much of the precision.

In analyses of black women, the MLE-based estimates of the RD were notably smaller than those from TMLE and its
We demonstrated the estimation of marginal RDs and risk ratios from weighted case–control data, using parametric methods and targeted semiparametric methods. Specifically, we used a substitution estimator based on MLE, TMLE and a rare outcomes refinement to assess the absolute and relative association between living in a high-poverty neighborhood and the risk of very preterm birth. Weights, based on known outcome prevalence, were used to account for the case–control sampling design.12

In this illustration, a relationship between living in a high-poverty neighborhood and risk of very preterm birth was observed among white women on both absolute and relative scales. Previous studies of neighborhood poverty and very preterm birth have not estimated absolute associations.28,29

The methods presented can be used to estimate parameters for other target populations. We focused on estimation of the marginal risk in each racial/ethnic group because the study samples were population-based and we were interested in estimating RRs and RDs specific to each of these groups. Other research questions could require averaging with respect to a different distribution of confounders. If a researcher were interested in a parameter for a subset of the study population (e.g., the exposed, one age group, or one income stratum), we would standardize (average) the conditional risk over the distribution of confounders for that subgroup. If a researcher were interested in estimating a parameter for an external population, confounder data from that other population would be used when generating the weighted mean of predicted probabilities. A growing body of literature addresses generalization or transportation of estimates to population beyond the study population.37

Alternative parametric methods of obtaining RDs from case–control data exist, but generally do not directly yield marginal estimates. For example, risks and their contrasts can be estimated by updating the intercept of parametric logistic regression using the crude risk in the target population or the ratio of case–control sampling fractions.8–11 This method can be mapped into marginal risk estimates by averaging with respect to the (weighted) covariate distribution; however, simulations have shown greater bias relative to MLE when the regression model is misspecified.15

There are several strengths and limitations to using these analytic approaches with case–control data. The ability to estimate both additive and relative associations from case–control data is a clear strength.1–3 These analytic approaches require knowledge of the population prevalence of the outcome, which is not necessarily assessed in a case–control design. However, an estimate may be available from an alternative data source, and a range of estimates compared if there is uncertainty.38

The two semiparametric methods (TMLE and the rare outcomes refinement) relax modeling assumptions and optimize control for confounders, although it is important to note that

| TABLE. Estimated Risks, Marginal Risk Difference, and Marginal Risk Ratio for Very Preterm Birth Under Conditions of High- and Low-neighborhood Poverty, by Race/Ethnicity and Estimation Method |
|-------------------------------------------------------------|-------------------------------------------------------------|-------------------------------------------------------------|-------------------------------------------------------------|
| Risk Per 100 (95% CI) Under Exposure (High-neighborhood Poverty) | Risk Per 100 (95% CI) Under No Exposure (Low-neighborhood Poverty) | Marginal Risk Difference Per 100 (95% CI) | Marginal Risk Ratio (95% CI) |
| White | 1.0 (0.6, 1.4) | 0.4 (0.3, 0.6) | 0.6 (0.1, 1.1) | 2.4 (1.2, 4.6) |
| MLE | 1.0 (−0.1, 2.0) | 0.5 (−1.2, 2.1) | 0.5 (−1.1, 2.1) | 2.0 (0.8, 5.1) |
| TMLE | 1.0 (0.7, 1.3) | 0.5 (0.2, 0.8) | 0.5 (0.0, 1.0) | 2.0 (1.1, 3.6) |
| Rare† | 0.8 (0.6, 1.1) | 0.2 (−0.3, 0.6) | 1.2 (0.7, 2.0) |
| Hispanic | 1.0 (0.7, 1.3) | 0.9 (0.1, 1.7) | 0.2 (−0.8, 1.1) | 1.2 (0.7, 2.2) |
| MLE | 1.0 (0.2, 1.9) | 0.9 (0.5, 1.2) | 0.2 (−0.3, 0.7) | 1.2 (0.8, 1.9) |
| TMLE | 1.0 (0.7, 1.3) | 0.5 (0.0, 1.0) | 0.5 (0.1, 1.1) | 2.4 (1.2, 4.6) |
| Rare† | 0.5 (0.3, 0.7) | 0.3 (−0.1, 0.6) | 1.2 (0.6, 2.0) |
| Black | 2.9 (2.1, 3.7) | 2.4 (1.7, 3.1) | 0.5 (−1.0, 2.0) | 1.2 (0.7, 2.1) |
| MLE | 3.2 (1.5, 4.8) | 2.3 (1.5, 3.2) | 0.8 (−1.0, 2.7) | 1.4 (0.8, 2.2) |
| TMLE | 3.2 (2.3, 4.1) | 2.3 (1.7, 3.0) | 0.8 (−0.5, 2.1) | 1.4 (0.9, 2.1) |

95% CIs are based on bootstrap with 1000 iterations.
†Rare outcomes targeted minimum loss-based estimation.
MLE indicates maximum likelihood estimation; TMLE, targeted maximum likelihood estimation.
unmeasured confounders remain a concern. The semiparametric approaches are more computationally intensive, and estimates are typically more variable than those from a correctly specified parametric model. However, given that the correct model is not typically known, a semi-parametric approach is often warranted. Both the TMLE and rare outcomes refinement are double-robust, providing two chances (the outcome regression and exposure mechanism) to obtain a consistent estimate, and increased efficiency if both mechanisms are correctly specified. Although the rare outcomes methodology increases the stability and precision of the analysis, it requires that the conditional risk of the rare outcome be bounded from above. Cross validation can be used to select bounds if they are not known in a particular application.22

We have demonstrated how to estimate both additive and relative associations from case–control data using parametric and semiparametric targeted approaches. Furthermore, we provided R code for the application of weighted MLE, TMLE, and the rare outcomes refinement to facilitate implementation by other researchers (Supplemental Digital Content 1; http://links.lww.com/EDE/B34).17,39,40 Flexibility to estimate a wider range of parameters allows the scientific question to drive the analysis and has potential to improve knowledge gained from case–control designs.17,18

ACKNOWLEDGMENTS

The authors wish to thank Mark van der Laan for reviewing the manuscript; Martin Kharrazi for providing access to the California Very Preterm Birth study; and Juan Yang, Craig Wolff, Steve Graham, and Charlene Sacramento for assistance with geocoding and data management. Data from prenatal screening were provided by the Genetic Disease Screening Program, and birth records were provided by the Health Information and Research Section, California Department of Public Health. All analyses, interpretations, and conclusions reached are those of the authors and not the California Department of Public Health.

REFERENCES