Chapter 13
Independent Case-Control Studies

Sherri Rose, Mark J. van der Laan

Case-control study designs are frequently used in public health and medical research to assess potential risk factors for disease. These study designs are particularly attractive to investigators researching rare diseases, as they are able to sample known cases of disease vs. following a large number of subjects and waiting for disease onset in a relatively small number of individuals.

Case-control sampling is a biased design. Bias occurs due to the disproportionate number of cases in the sample vs. the population.

Researchers commonly employ the use of logistic regression in a parametric statistical model, ignoring the biased design, and estimate the conditional odds ratio of having disease given the exposure of interest $A$ and measured covariates $W$.

Our proposed case-control-weighted TMLE for case-control studies relies on knowledge of the true prevalence probability, or a reasonable estimate of this probability, to eliminate the bias of the case-control sampling design. We use the prevalence probability in case-control weights, and our case-control weighting scheme successfully maps the TMLE for a random sample into a method for case-control sampling. The case-control-weighted TMLE (CCW-TMLE) is an efficient estimator for the case-control sample when the TMLE for the random sample is efficient. In addition, the CCW-TMLE inherits the robustness properties of the TMLE for the random sample.

13.1 Data, Model, and Target Parameter

Let us define a simple example with $X = (W, A, Y) \sim P_{X,0}$ as the full-data experimental unit and corresponding distribution $P_{X,0}$ of interest, which consists of baseline covariates $W$, exposure variable $A$, and a binary outcome $Y$ that defines case or...
control status. In previous chapters, our target parameter of interest was the causal risk difference, which we now denote

$$\psi_{RD,0}^F = \psi_{RD}^F(X,0) = E_{X,0}(Y | A = 1, W) - E_{X,0}(Y | A = 0, W)$$

$$= E_{X,0}(Y_1) - E_{X,0}(Y_0)$$

$$= P_{X,0}(Y_1 = 1) - P_{X,0}(Y_0 = 1)$$

for binary $A$, binary $Y$, and counterfactual outcomes $Y_0$ and $Y_1$, where $F$ indicates “full data.” Other common parameters of interest include the causal relative risk and the causal odds ratio, given by

$$\psi_{RR,0}^F = \frac{P_{X,0}(Y_1 = 1)}{P_{X,0}(Y_0 = 1)}$$

and

$$\psi_{OR,0}^F = \frac{P_{X,0}(Y_1 = 1)P_{X,0}(Y_0 = 0)}{P_{X,0}(Y_1 = 0)P_{X,0}(Y_0 = 1)}.$$  

We describe the case-control design as first sampling $(W_1, A_1)$ from the conditional distribution of $(W, A)$, given $Y = 1$ for a case. One then samples $J$ controls $(W_0^j, A_0^j)$ from $(W, A)$, given $Y = 0, j = 1, \ldots, J$. The observed data structure in independent case-control sampling is then defined by

$$O = \{(W_1, A_1), (W_0^j, A_0^j : j = 1, \ldots, J)\} \sim P_0,$$  

with