On variance of the treatment effect in the treated using inverse probability weighting

Sarah A. Reifeis[1,*] and Michael G. Hudgens[1,**]

[1]*Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA*

[*]*sreifeis@email.unc.edu*

[**]*mhudgens@email.unc.edu*

Contents

Web Appendix 1

Variance Estimators of the ATT estimator

Assume we observe n independent and identically distributed (i.i.d.) copies of (L,A,Y) denoted by (L_i,A_i,Y_i) for $i=1,...,n$. Evaluating the IPW ATT estimator $\widehat{ATT}$ in equation 1 of the main text first entails estimating the inverse probability weights $\hat{W}_i$ for $i=1,...,n$. The weights are estimated by fitting the logistic regression model

$$\text{logit}\{P(A=1|L)\} = \alpha_0 + \alpha_1^T L \tag{1}$$

where L represents a (column) vector of J measured pre-exposure variables and α_1 is a parameter vector of length J. Let $\alpha = (\alpha_0, \alpha_1^T)$ and let $\hat{\alpha}$ be the maximum likelihood estimator (MLE) of α obtained by fitting model 1. Letting $h(L;\alpha) = P(A=1|L)/P(A=0|L) = \exp(\alpha_0 + \alpha_1^T L)$, the estimated weight for individual i is given by $\hat{W}_i = W(A_i,L_i;\hat{\alpha}) = A_i + (1-A_i)h(L_i;\hat{\alpha})$. Below the large sample properties of $\widehat{ATT}$ are considered under the following identification conditions: stable unit treatment value assumption [1]; positivity [2], i.e., $P(A=0|L=l) > 0$ for all l where $dF_L(l) > 0$ and F_L is the CDF of L given $A=1$; partial conditional exchangeability [3], i.e., $Y^0 \perp\!\!\!\perp A|L$; and correct specification of the model for $A|L$.

The asymptotic distribution of the IPW ATT estimator in equation 1 of the main text can be derived using standard estimating equation theory. In particular, let

$$\psi(Y_i,A_i,L_i,\alpha,\mu) = \begin{pmatrix} \psi_\alpha(A_i,L_i,\alpha) \\ \psi_1(Y_i,A_i,L_i,\alpha,\mu) \\ \psi_0(Y_i,A_i,L_i,\alpha,\mu) \end{pmatrix} = \begin{pmatrix} \{A_i - e(L_i;\alpha)\}(1,L_i^T)^T \\ W(A_i,L_i;\alpha)A_i(Y_i-\mu_1) \\ W(A_i,L_i;\alpha)(1-A_i)(Y_i-\mu_0) \end{pmatrix}$$

where $\mu = (\mu_1,\mu_0)$, ψ_α denotes the $J+1$ vector of score functions from the log likelihood corresponding to model 1 above, $e(L_i;\alpha) = P(A_i=1|L_i) = h(L_i;\alpha)/\{1+h(L_i;\alpha)\}$ is the propensity score, and $W_i = W(A_i,L_i;\alpha) = A_i + (1-A_i)h(L_i;\alpha)$ [4]. The functions ψ_1 and ψ_0 correspond to the first and second ratios of the ATT estimator in equation 1, respectively.

Let $\xi = (\alpha_0, \alpha_1^T, \mu_1, \mu_0)^T$ and let $\hat{\xi} = (\hat{\alpha}_0, \hat{\alpha}_1^T, \hat{\mu}_1, \hat{\mu}_0)^T$, where $\hat{\xi}$ solves the estimating equations $\sum_i \psi(Y_i,A_i,L_i,\alpha,\mu) = 0$ and $\widehat{ATT} = \hat{\mu}_1 - \hat{\mu}_0$ is the ATT estimator in equation 1 of the main text. Then under suitable regularity conditions [5]

$$\sqrt{n}(\hat{\xi} - \xi) \to^d N(0, V(\xi))$$

where $V(\xi) = \mathbb{A}(\xi)^{-1}\mathbb{B}(\xi)\{\mathbb{A}(\xi)^{-1}\}^T$ with $\mathbb{B}(\xi) = E\{\psi(Y,A,L,\xi)\psi(Y,A,L,\xi)^T\}$, $\mathbb{A}(\xi) = E\{-\dot{\psi}(Y,A,L,\xi)\}$, and $\dot{\psi}(Y,A,L,\xi) = \partial\psi(Y,A,L,\xi)/\partial\xi^T$. It follows from the Delta method that $\widehat{ATT}$ is consistent and asymptotically normal, i.e.,

$$\sqrt{n}\left(\widehat{ATT} - ATT\right) \xrightarrow{d} N(0,\Sigma).$$

where $\Sigma = \nabla g(\xi)^T V(\xi) \nabla g(\xi)$ is the asymptotic variance of the ATT estimator, $\nabla g(\xi)^T = (0, 0_J^T, 1, -1)$, and 0_J is the 0 vector of length J. Let $\hat{\Sigma}$ denote a consistent estimator for Σ obtained by substituting $\hat{V}(\hat{\xi})$ for $V(\xi)$ where the expectations in V are replaced by their empirical counterparts and $\hat{\xi}$ is substituted for ξ. Then in large samples the variance of $\widehat{ATT}$ can be approximated by $\hat{\Sigma}/n$, the stacked estimating equations (SEE) variance estimator.

Now suppose instead that the weights W_i are assumed known and need not be estimated. Let $\widehat{ATT}^*$ denote the estimator in equation 1 from the main text where $\hat{W}_i$ is replaced with W_i. Then $\widehat{ATT}^*$ is consistent for ATT and asymptotically normal with asymptotic variance

$$\Sigma^* = p_1^{-2}\left[E\{A_i(Y_i^1 - \mu_1)^2\} + E\{(Y_i^0 - \mu_0)^2 h(L_i;\alpha)e(L_i;\alpha)\}\right]. \tag{2}$$

where $p_1 = P(A = 1)$, as shown in the next section of Web Appendix 1. Let $\hat{\Sigma}^*$ represent an estimator for Σ^* obtained by substituting p_1, μ_1, and the expectations in equation 2 with their empirical counterparts, where α is assumed known. Then $\hat{\Sigma}^*/n$ denotes the naïve variance estimator discussed in the Methods section of the main text. In Web Appendix 1 it is also shown that

$$\Sigma = \Sigma^* + p_1^{-2}(c_{11} + c_{22} - 2c_{12}) \tag{3}$$

and the explicit forms of c_{km}, $k,m \in \{1,2\}$ are given. In general, the sign of the second term on the right side of the equality of equation 3 can be either positive or negative.

Derivation and Relationship of Σ and Σ^*

Continuing with the case introduced above where the weights $W_i = W(A_i, L_i; \alpha)$ are assumed known, let

$$\psi^*(Y_i, A_i, L_i, \mu) = \begin{pmatrix} \psi_1^*(Y_i, A_i, L_i, \mu) \\ \psi_0^*(Y_i, A_i, L_i, \mu) \end{pmatrix} = \begin{pmatrix} W_i A_i (Y_i - \mu_1) \\ W_i (1 - A_i)(Y_i - \mu_0) \end{pmatrix}$$

where μ is as defined previously. Let $\hat{\mu}^* = (\hat{\mu}_1^*, \hat{\mu}_0^*)$ where $\hat{\mu}^*$ solves the estimating equations $\Sigma_i \psi^*(Y_i, A_i, L_i, \mu) = 0$ and $\widehat{ATT}^* = \hat{\mu}_1^* - \hat{\mu}_0^*$. Following the same logic used for Σ above, $\Sigma^* = \left(1, -1\right) V^*(\mu) \left(1, -1\right)^T$ where $V^*(\mu) = \mathbb{A}^*(\mu)^{-1} \mathbb{B}^*(\mu) \{\mathbb{A}^*(\mu)^{-1}\}^T$ and

$$\begin{aligned}
\mathbb{A}^*(\mu) &= -E \begin{pmatrix} \partial \psi_1^* / \partial \mu_1 & \partial \psi_1^* / \partial \mu_0 \\ \partial \psi_0^* / \partial \mu_1 & \partial \psi_0^* / \partial \mu_0 \end{pmatrix} \\
&= \begin{pmatrix} E[A_i] & 0 \\ 0 & E\left[(1 - A_i) \frac{P(A_i=1|L_i)}{P(A_i=0|L_i)}\right] \end{pmatrix} \\
&= \begin{pmatrix} P(A=1) & 0 \\ 0 & E_L\left[E_{A|L}\left[(1 - A_i)\right] \frac{P(A_i=1|L_i)}{P(A_i=0|L_i)}\right] \end{pmatrix} \\
&= p_1 I_2
\end{aligned}$$

where I_2 denotes a 2×2 identity matrix, and

$$\begin{aligned}
\mathbb{B}^*(\mu) &= E \begin{pmatrix} \psi_1^{*2} & \psi_1^* \psi_0^* \\ \psi_0^* \psi_1^* & \psi_0^{*2} \end{pmatrix} \\
&= \begin{pmatrix} E[A_i (Y_i - \mu_1)^2] & 0 \\ 0 & E\left[(1 - A_i) \frac{(Y_i - \mu_0)^2 (P(A_i=1|L_i))^2}{(P(A_i=0|L_i))^2}\right] \end{pmatrix} \\
&= \begin{pmatrix} E[A_i (Y_i^1 - \mu_1)^2] & 0 \\ 0 & E\left[(Y_i^0 - \mu_0)^2 h(L_i; \alpha) e(L_i; \alpha)\right] \end{pmatrix}
\end{aligned}$$

Then $V^*(\mu) = p_1^{-2} \mathbb{B}^*(\mu)$, which results in the expression for Σ^* from equation 2.

Returning to the usual case where the weights W_i are unknown and need to be estimated, again consider the vector of estimating functions $\psi(Y_i, A_i, L_i, \alpha, \mu)$ defined in Web Appendix 1 above. Using block notation, the components of $V(\xi)$ may be expressed as

$$\mathbb{A}(\xi) = -E \left(\begin{array}{cc|cc} \partial \psi_{\alpha_0} / \partial \alpha_0 & \partial \psi_{\alpha_0} / \partial \alpha_1 & \partial \psi_{\alpha_0} / \partial \mu_1 & \partial \psi_{\alpha_0} / \partial \mu_0 \\ \partial \psi_{\alpha_1} / \partial \alpha_0 & \partial \psi_{\alpha_1} / \partial \alpha_1 & \partial \psi_{\alpha_1} / \partial \mu_1 & \partial \psi_{\alpha_1} / \partial \mu_0 \\ \hline \partial \psi_1 / \partial \alpha_0 & \partial \psi_1 / \partial \alpha_1 & \partial \psi_1 / \partial \mu_1 & \partial \psi_1 / \partial \mu_0 \\ \partial \psi_0 / \partial \alpha_0 & \partial \psi_0 / \partial \alpha_1 & \partial \psi_0 / \partial \mu_1 & \partial \psi_0 / \partial \mu_0 \end{array}\right) = \begin{pmatrix} a_{11} & 0_{(J+1) \times 2} \\ a_{21} & \mathbb{A}^*(\mu) \end{pmatrix}$$

$$\mathbb{B}(\xi) = E \left(\begin{array}{cc|cc} \psi_{\alpha_0}^2 & \psi_{\alpha_0} \psi_{\alpha_1}^T & \psi_{\alpha_0} \psi_1 & \psi_{\alpha_0} \psi_0 \\ \psi_{\alpha_1} \psi_{\alpha_0} & \psi_{\alpha_1} \psi_{\alpha_1}^T & \psi_{\alpha_1} \psi_1 & \psi_{\alpha_1} \psi_0 \\ \hline \psi_1 \psi_{\alpha_0} & \psi_1 \psi_{\alpha_1}^T & \psi_1^2 & \psi_1 \psi_0 \\ \psi_0 \psi_{\alpha_0} & \psi_0 \psi_{\alpha_1}^T & \psi_0 \psi_1 & \psi_0^2 \end{array}\right) = \begin{pmatrix} b_{11} & b_{21}^T \\ b_{21} & b_{22} \end{pmatrix}$$

where $\psi_{\alpha_0}, \psi_{\alpha_1}$ correspond to the score functions for the intercept and the J covariates, respectively, from the logistic regression model 1 in Web Appendix 1, and in general $0_{m\times n}$ denotes an $m \times n$ zero matrix and 0_m denotes a column vector of m zeros. Note $b_{22} = \mathbb{B}^*(\mu)$ for all $\alpha \in \mathbb{R}^{J+1}$, and recall that $\mathbb{A}^*(\mu) = p_1 I_2$.

Next note

$$\mathbb{A}(\xi)^{-1} = \begin{pmatrix} a_{11}^{-1} & 0_{(J+1)\times 2} \\ -p_1^{-1} a_{21} a_{11}^{-1} & p_1^{-1} I_2 \end{pmatrix}$$

and $a_{11} = b_{11}$ [6, Lemma 7.3.11], implying

$$\begin{aligned} V(\xi) &= \begin{pmatrix} a_{11}^{-1} & 0_{(J+1)\times 2} \\ -p_1^{-1} a_{21} a_{11}^{-1} & p_1^{-1} I_2 \end{pmatrix} \begin{pmatrix} b_{11} & b_{21}^T \\ b_{21} & b_{22} \end{pmatrix} \begin{pmatrix} a_{11}^{-1} & -p_1^{-1} a_{11}^{-1} a_{21}^T \\ 0_{(J+1)x2} & p_1^{-1} I_2 \end{pmatrix} \\ &= \begin{pmatrix} a_{11}^{-1} & p_1^{-1} a_{11}^{-1} (-a_{21} + b_{21})^T \\ p_1^{-1}(-a_{21} + b_{21}) a_{11}^{-1} & p_1^{-2}\{(a_{21} - b_{21}) a_{11}^{-1} (a_{21} - b_{21})^T - b_{21} a_{11}^{-1} b_{21}^T + b_{22}\} \end{pmatrix} \end{aligned}$$

By the Delta method, $\Sigma = \nabla g(\xi)^T V(\xi) \nabla g(\xi)$ where $\nabla g(\xi)^T = (0, 0_J^T, 1, -1)$. Let c be the 2×2 matrix $c = (a_{21} - b_{21}) a_{11}^{-1} (a_{21} - b_{21})^T - b_{21} a_{11}^{-1} b_{21}^T$, with elements $c_{11}, c_{12}, c_{21}, c_{22}$. Then

$$\begin{aligned} \Sigma &= \nabla g(\xi)^T \begin{pmatrix} a_{11}^{-1} & p_1^{-1} a_{11}^{-1} (-a_{21} + b_{21})^T \\ p_1^{-1}(-a_{21} + b_{21}) a_{11}^{-1} & p_1^{-2}(c + b_{22}) \end{pmatrix} \nabla g(\xi) \\ &= p_1^{-2} \left[E\{A_i (Y_i^1 - \mu_1)^2\} + E\{(Y_i^0 - \mu_0)^2 h(L_i;\alpha) e(L_i;\alpha)\} + c_{11} + c_{22} - c_{12} - c_{21} \right] \\ &= \Sigma^* + p_1^{-2}(c_{11} + c_{22} - 2c_{12}) \end{aligned}$$

where the last equality follows from the derivation above of Σ^*, and $c_{12} = c_{21}$ because $V(\xi)$ is symmetric.

Next note

$$a_{21} = \begin{pmatrix} 0 & 0_J^T \\ -E\{(1 - A_i)(Y_i - \mu_0) h(L_i;\alpha)\} & -E\{(1 - A_i)(Y_i - \mu_0) h(L_i;\alpha) L_i^T\} \end{pmatrix}$$

$$b_{21} = \begin{pmatrix} E\{A_i (Y_i - \mu_1)(1 - e(L_i;\alpha))\} & E\{A_i (Y_i - \mu_1)(1 - e(L_i;\alpha)) L_i^T\} \\ -E\{(1 - A_i)(Y_i - \mu_0) h(L_i;\alpha) e(L_i;\alpha)\} & -E\{(1 - A_i)(Y_i - \mu_0) h(L_i;\alpha) e(L_i;\alpha) L_i^T\} \end{pmatrix}$$

where $e(L_i;\alpha)$ and $h(L_i;\alpha)$ are as defined previously.

Assuming the propensity score model 1 from Web Appendix 1 and conditional exchangeability, the expectations above can be expressed

$$E\left[A_i(Y_i-\mu_1)\{1-e(L_i;\alpha)\}(1,L_i^T)\right] = E\left[A_i(Y_i^1-\mu_1)\{1-P(A_i=1|L_i)\}(1,L_i^T)\right]$$
$$= E_L\left[E_{Y^1|L}(Y_i^1-\mu_1)\ P(A_i=1|L_i)\{1-P(A_i=1|L_i)\}\ (1,L_i^T)\right]$$

and similarly $E\{(1-A_i)(Y_i-\mu_0)h(L_i;\alpha)(1,L_i^T)\} = E_L\{E_{Y^0|L}(Y_i^0-\mu_0)\ P(A_i=1|L_i)(1,L_i^T)\}$ and $E\{(1-A_i)(Y_i-\mu_0)e(L_i;\alpha)h(L_i;\alpha)(1,L_i^T)\} = E_L\{E_{Y^0|L}(Y_i^0-\mu_0)\ P(A_i=1|L_i)^2(1,L_i^T)\}$.

Likewise, a_{11} can be written

$$a_{11} = \begin{pmatrix} E(\psi_{\alpha_0}^2) & E(\psi_{\alpha_0}\psi_{\alpha_1})^T \\ E(\psi_{\alpha_0}\psi_{\alpha_1}) & E(\psi_{\alpha_1}\psi_{\alpha_1}^T) \end{pmatrix}$$

where

$$E(\psi_{\alpha_0}^2) = E\{A_i - 2A_iP(A_i=1|L_i) + P(A_i=1|L_i)^2\}$$
$$= E_L[P(A_i=1|L_i)\{1-P(A_i=1|L_i)\}]$$

with similar derivations for $E(\psi_{\alpha_0}\psi_{\alpha_1}) = E_L[P(A_i=1|L_i)\{1-P(A_i=1|L_i)\}L_i]$ and $E(\psi_{\alpha_1}\psi_{\alpha_1}^T) = E_L[P(A_i=1|L_i)\{1-P(A_i=1|L_i)\}L_iL_i^T]$.

Using the results above, explicit values for each element of the c matrix can be calculated for given distributions of L, $A|L$, $Y^0|L$, and $Y^1|L$. This is demonstrated in the Asymptotic Calculations section of the main text for four example scenarios. The R code used for these calculations is included below in Web Appendix 2.

Expected value of ATT weights

The expected value of the weights proposed by Sato and Matsuyama [4] equals

$$E[W_i] = E_{A,L}\left[A_i + (1-A_i)\frac{P(A_i=1|L_i)}{P(A_i=0|L_i)}\right]$$
$$= p_1 + E_L\left[E_{A|L}(1-A_i)\frac{P(A_i=1|L_i)}{P(A_i=0|L_i)}\right]$$
$$= p_1 + E_L\left[P(A_i=1|L_i)\right] = 2p_1$$

Web Appendix 2

The code below was written for the R environment in R version 3.6.3 [7].

Asymptotic Calculations

First set the values of the parameters from scenario (i) in the main text.

```
EL <- 0.5 ; a0 <- -1 ; a1 <- -2
ba <- -1 ; bL <- -1.5 ; baL <- 1.5 ; sdY <- 0.5
```

From these defined values we can solve for the other needed quantities.

```
EA_L1 <- exp(a0 + a1) / (1 + exp(a0 + a1))
EA_L0 <- exp(a0) / (1 + exp(a0))

EY1_L1 <- ba + bL + baL #no intercept term
EY1_L0 <- ba
EY0_L1 <- bL
EY0_L0 <- 0
VarY0_L <- sdY^2
VarY1_L <- sdY^2

EA <- EA_L0*(1-EL) + EA_L1*(EL)
EL_A1 <- (1/EA) * EA_L1 * EL
mu0 <- bL * EL_A1
mu1 <- ba + bL*EL_A1 + baL*EL_A1
ATT <- mu1-mu0
```

These values can be plugged in to calculate the elements of the a_{21}, b_{21}, and a_{11}^{-1} matrices.

```
## Calculate required expectations for (a21 - b21),
##  b21, and a11^{-1} matrices
a21_b21.1 <- (EY1_L0 - mu1)*EA_L0*(1-EA_L0)*(1-EL) +
                (EY1_L1 - mu1)*EA_L1*(1-EA_L1)*EL
a21_b21.2 <- (EY0_L0 - mu0)*EA_L0*(1-EA_L0)*(1-EL) +
                (EY0_L1 - mu0)*EA_L1*(1-EA_L1)*EL
a21_b21.3 <- (EY1_L0 - mu1)*EA_L0*(1-EA_L0)*0*(1-EL) +
                (EY1_L1 - mu1)*EA_L1*(1-EA_L1)*1*EL
a21_b21.4 <- (EY0_L0 - mu0)*EA_L0*(1-EA_L0)*0*(1-EL) +
                (EY0_L1 - mu0)*EA_L1*(1-EA_L1)*1*EL

a21_b21 <- matrix(c(-a21_b21.1, -a21_b21.2, -a21_b21.3, -a21_b21.4),
                    nrow=2, ncol=2)

b21.2 <- (EY0_L0 - mu0)*(EA_L0^2)*(1-EL) +
              (EY0_L1 - mu0)*(EA_L1^2)*EL
b21.4 <- (EY0_L0 - mu0)*(EA_L0^2)*0*(1-EL) +
              (EY0_L1 - mu0)*(EA_L1^2)*1*EL

b21 <- matrix(c(a21_b21.1, -b21.2, a21_b21.3, -b21.4),
                nrow=2, ncol=2)

a11_1 <- EA_L0*(1-EA_L0)*(1-EL) + EA_L1*(1-EA_L1)*EL
a11_2 <- EA_L0*(1-EA_L0)*0*(1-EL) + EA_L1*(1-EA_L1)*1*EL
a11_3 <- EA_L0*(1-EA_L0)*(0^2)*(1-EL) + EA_L1*(1-EA_L1)*(1^2)*EL

a11 <- matrix(c(a11_1, a11_2, a11_2, a11_3),
                nrow=2, ncol=2)
a11_inv <- solve(a11)
```

What remains is simply using matrix algebra to calculate values of the constant and Σ^*.

```r
## Calculate constant
c <- a21_b21 %*% a11_inv %*% t(a21_b21) - b21 %*% a11_inv %*% t(b21)
c_scaled <- (1/EA^2)*c  # (1/P(A=1)^2) * c
gg <- cbind(c(1, -1))

constant <- t(gg) %*% c_scaled %*% gg

## Calculate Sigma* and Sigma
EY0_mu02_L0 <- (VarY0_L + EY0_L0^2) - 2*mu0*EY0_L0 + mu0^2
EY0_mu02_L1 <- (VarY0_L + EY0_L1^2) - 2*mu0*EY0_L1 + mu0^2
EY1_mu12_L0 <- (VarY1_L + EY1_L0^2) - 2*mu1*EY1_L0 + mu1^2
EY1_mu12_L1 <- (VarY1_L + EY1_L1^2) - 2*mu1*EY1_L1 + mu1^2

b22_1 <- (EA_L0*EY1_mu12_L0)*(1-EL) + (EA_L1*EY1_mu12_L1)*EL
b22_2 <- ((EA_L0^2/(1-EA_L0))*EY0_mu02_L0)*(1-EL) +
            ((EA_L1^2/(1-EA_L1))*EY0_mu02_L1)*EL

Sig_star <- (b22_1 + b22_2)/(EA^2)
Sig <- Sig_star + constant
SD_ratio <- sqrt(Sig)/sqrt(Sig_star)

df <- data.frame(cbind(ATT, constant, Sig_star, Sig, SD_ratio))
colnames(df) <- c("ATT", "Constant", "Sigma^*", "Sigma", "SD Ratio")
print(df)
```

```
##          ATT Constant  Sigma^*    Sigma  SD Ratio
## 1 -0.7751385 1.635956 2.263171 3.899128  1.312578
```

These results are presented in the first row of Table 2.

Simulated Data Analysis

Using the population parameters defined above, we can simulate an example data set of 1000 individuals. After generating L, A, and Y, the ATT weights are computed as in the main text.

```
set.seed(42)
n <- 1000

L <- rbinom(n, 1, prob = EL)
lp <- exp(a0 + a1*L)
A <- rbinom(n, size = 1, prob = lp/(1+lp))
Y <- rnorm(n, mean = ba*A + bL*L + baL*A*L, sd = sdY)

psmod <- glm(A ~ L, family = binomial(link = "logit"))
wt.att <- ifelse(A == 0, exp(psmod$linear.predictors), 1)

dat <- data.frame(cbind(L, A, Y, wt.att))
```

The following are helper functions defined for use within the geex function `m_estimate`, which will allow us to compute the standard errors for the stacked estimating equations (SEE) and naïve variance estimators.

```
estfun <- function(data, model){
  L <- model.matrix(model, data=data)
  A <- model.response(model.frame(model, data=data))
  Y <- data$Y

  function(theta){
    p  <- length(theta)
    p1 <- length(coef(model))
    lp  <- L %*% theta[1:p1]
    rho <- plogis(lp)
```

```
    IPW <- ifelse(A == 1, 1, exp(lp))

    score_eqns <- apply(L, 2, function(x) sum((A - rho) * x))
    ce1 <- IPW*(A==1)*(Y - theta[p-1])
    ce0 <- IPW*(A==0)*(Y - theta[p])

    c(score_eqns,
      ce1,
      ce0)
  }
}

estfun_nolr <- function(data){
  A <- data$A
  Y <- data$Y
  IPW <- data$wt.att

  function(theta){
    ce1 <- IPW*(A==1)*(Y - theta[1])
    ce0 <- IPW*(A==0)*(Y - theta[2])

    c(ce1,
      ce0)
  }
}
```

Fitting the weighted linear regression model yields the estimated counterfactual means, from which we can compute the estimated ATT.

```
fit <- geeglm(Y ~ A, data = dat, std.err = 'san.se',
              weights = wt.att, id=1:nrow(dat),
              corstr="independence")
```

```r
mu1_hat <- mean(fit$fitted.values[fit$dat$A==1])
mu0_hat <- mean(fit$fitted.values[fit$dat$A==0])

ATT_Est <- fit$coefficients[2]  # = mu1_hat - mu0_hat
```

Finally, the geex package is used to estimate the SEs of the estimated ATT using both the SEE and the naïve estimators. The naïve SEs are also computed with the `geeglm` function to check the output from `m_estimate`.

```r
## Accounting for weight estimation
results <- m_estimate(
               estFUN = estfun,
               data  = dat,
               roots = c(coef(psmod), mu1_hat, mu0_hat),
               compute_roots = FALSE,
               outer_args = list(model = psmod))

## b22 + [1/P(A=1)^2]c
vcov_sEE <- vcov(results)[3:4, 3:4]

## Assuming weights are known
results_nolr <- m_estimate(
                    estFUN = estfun_nolr,
                    data  = dat,
                    roots = c(mu1_hat, mu0_hat),
                    compute_roots = FALSE)

## b22
vcov_GEE <- vcov(results_nolr)

## Naive Variance from geeglm for comparison
vcov_geeglm <- (summary(fit)$coefficients[2,2])^2
```

```
Sig_est <- t(gg) %*% vcov_sEE %*% gg
Sig_star_est <- t(gg) %*% vcov_GEE %*% gg

df <- data.frame(cbind(ATT_Est, sqrt(Sig_est),
                       sqrt(Sig_star_est), sqrt(vcov_geeglm)))
colnames(df) <- c("Est ATT", "Est SEE SE",
                  "Est Naive SE (geex)", "Est Naive SE (geeglm)")
print(df)
```

```
##      Est ATT Est SEE SE Est Naive SE (geex) Est Naive SE (geeglm)
## A -0.7543794 0.05830972          0.04407246             0.04407246
```

The SE estimates from `geeglm` and from `geex` when weights are assumed known are the same. All estimates resemble the results presented in Table 3, but do not match exactly since this code was only run on one example data set and the Table 3 results are averaged over 1000 data sets.

Note that when performing the analysis for a large number of simulated data sets or, e.g., a large genomics data set such as METSIM with hundreds or thousands of individuals and outcomes, there may be a practical need to run the code for analyzing these data sets simultaneously on a computing cluster.

Web Appendix 3

Varied Sample Size Results for Simulation Study Scenarios (i)-(iv)

The main simulation study was repeated with sample sizes $n = 500$ and $n = 2000$. The results, given in Web Table 1 below, are similar to those reported in Table 3. For scenario (iv) the SEE estimator tends to underestimate the variability of the ATT estimator for the sample sizes considered, resulting in confidence interval coverage slightly below the nominal level. We note

however that the relative bias of the SEE estimator decreases as the sample size increases (0.12 for $n = 500$, 0.10 for $n = 1000$, and 0.09 for $n = 2000$). The empirical sandwich estimator has been shown in other settings to underestimate the true variance when the sample size is small [8, 9, 10], in which case bias-corrected variance estimators might be considered.

Web Table 1: Empirical standard error, average estimated standard error using the SEE and naïve variance estimates, 95% confidence interval coverage, and $\widehat{ASE}$ ratio (SEE/Naïve) with sample sizes of 500 and 2000 for each simulated scenario.

			SEE		Naïve		
Scenario	n	ESE	$\widehat{ASE}$	Coverage	$\widehat{ASE}$	Coverage	$\widehat{ASE}$ Ratio
(i)	500	0.09	0.09	0.95	0.07	0.87	1.31
	2000	0.04	0.04	0.95	0.03	0.86	1.31
(ii)	500	0.05	0.05	0.95	0.09	1.00	0.56
	2000	0.03	0.03	0.96	0.05	1.00	0.56
(iii)	500	0.09	0.09	0.95	0.09	0.92	1.10
	2000	0.05	0.05	0.96	0.04	0.94	1.10
(iv)	500	0.15	0.14	0.93	0.21	1.00	0.64
	2000	0.08	0.07	0.93	0.11	0.99	0.65

$\widehat{ASE}$ = average estimated standard error; ESE = empirical standard error; SEE = stacked estimating equations.

Bootstrap SE Results

Simulation Study Scenarios (i)-(iv)

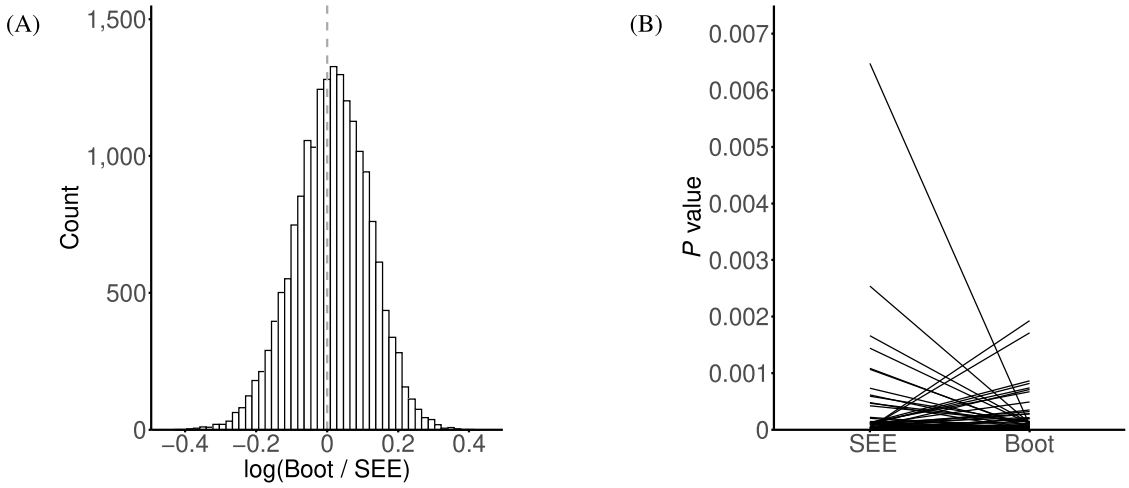
The main simulation study and METSIM data analysis were repeated using the bootstrap to estimate the SE. For each data set, the bootstrap SE estimate was computed using the following steps: (1) a simple random sample with replacement of size 1000 was drawn from the original data; (2) $\widehat{ATT}$ was computed; (3) steps (1) and (2) were repeated $B = 50$ times, and then the standard deviation of the $B = 50$ estimates of the ATT was computed. Results of the simulation study are presented in Web Table 2 and the METSIM data analysis results are displayed in Web Figure 1.

Web Table 2: Empirical standard error, average estimated standard error using bootstrap variance estimates, and 95% confidence interval coverage for each simulated scenario.

		Bootstrap	
Scenario	ESE	$\widehat{ASE}$	Coverage
(i)	0.06	0.06	0.95
(ii)	0.04	0.04	0.94
(iii)	0.07	0.07	0.95
(iv)	0.12	0.10	0.92

$\widehat{ASE}$ = average estimated standard error; ESE = empirical standard error.

METSIM Analysis



Web Figure 1: (A) Ratio of estimated standard errors (SEs) computed using the bootstrap estimator ($B = 50$) and $\hat{\Sigma}$, for the average treatment effect in the treated of each gene in the Metabolic Syndrome in Men (METSIM) data analysis. Vertical dashed line at zero denotes equality of the two SE estimates. (B) P-values (unadjusted) for both methods of SE estimation, for each of the top 50 genes as ranked by either method (66 genes depicted in total). Boot = bootstrap; SEE = stacked estimating equations.

Risk Ratio Simulations

The causal risk ratio in the treated is defined as μ_1/μ_0, and can be consistently estimated by the ratio of Hajek estimators from equation 2 of the main text. Two binary outcome scenarios, (v) and (vi), are defined in Web Table 3 and the corresponding asymptotic calculations are included in Web Table 4. The asymptotic calculations were determined as in Web Appendix 1, except that $\nabla g(\xi)^T = (0, 0_J^T, 1/\mu_0, -\mu_1/\mu_0^2)$. A simulation study of 1000 data sets with n=1000 was conducted for each scenario, and the results are presented in Web Table 5. The ratio estimator variance is consistently estimated using SEE in both cases, with the $\widehat{ASE}$ closely approximating the ESE and Wald CIs achieving nominal coverage. On the other hand, the naïve variance estimator over- and under-estimates the ratio estimator variance in scenarios (v) and (vi), respectively, resulting in Wald CIs that are conservative and anti-conservative.

Web Table 3: Distribution of L, exposure A, and potential outcome Y^a in two different scenarios for the CRRT, along with the marginal probability of exposure and the CRRT.

Scenario	L	$P(A=1\|L=l)$	$P(Y^a=1\|L=l)$	p_1	CRRT
(v)	Bern(0.2)	$0.6+0.2\,l$	$0.35+0.6\,l$	0.64	1
(vi)	Bern(0.5)	$0.4+0.2\,l$	$0.95-0.84\,a-0.65\,l+1.4\,a\,l$	0.5	1

Bern(π) = Bernoulli distribution with expectation π; CRRT = causal risk ratio in the treated; p_1 = marginal probability of exposure.

Web Table 4: The asymptotic variance of the CRRT estimator when weights are unknown (Σ) and known (Σ^*), and the ratio (Unknown / Known) of the asymptotic standard deviations.

Scenario	Σ	Σ^*	SD Ratio
(v)	3.04	4.88	0.79
(vi)	5.00	3.50	1.19

CRRT = causal risk ratio in the treated; SD = stndard deviation.

Web Table 5: Empirical standard error, average estimated standard error using the SEE and naïve variance estimates of the causal risk ratio in the treated estimator, 95% confidence interval coverage, and $\widehat{ASE}$ ratio (SEE/Naïve) for two simulated scenarios.

		SEE		Naïve		
Scenario	ESE	$\widehat{ASE}$	Coverage	$\widehat{ASE}$	Coverage	$\widehat{ASE}$ Ratio
(v)	0.06	0.06	0.94	0.07	0.99	0.79
(vi)	0.07	0.07	0.95	0.06	0.91	1.19

$\widehat{ASE}$ = average estimated standard error; ESE = empirical standard error; SEE = stacked estimating equations.

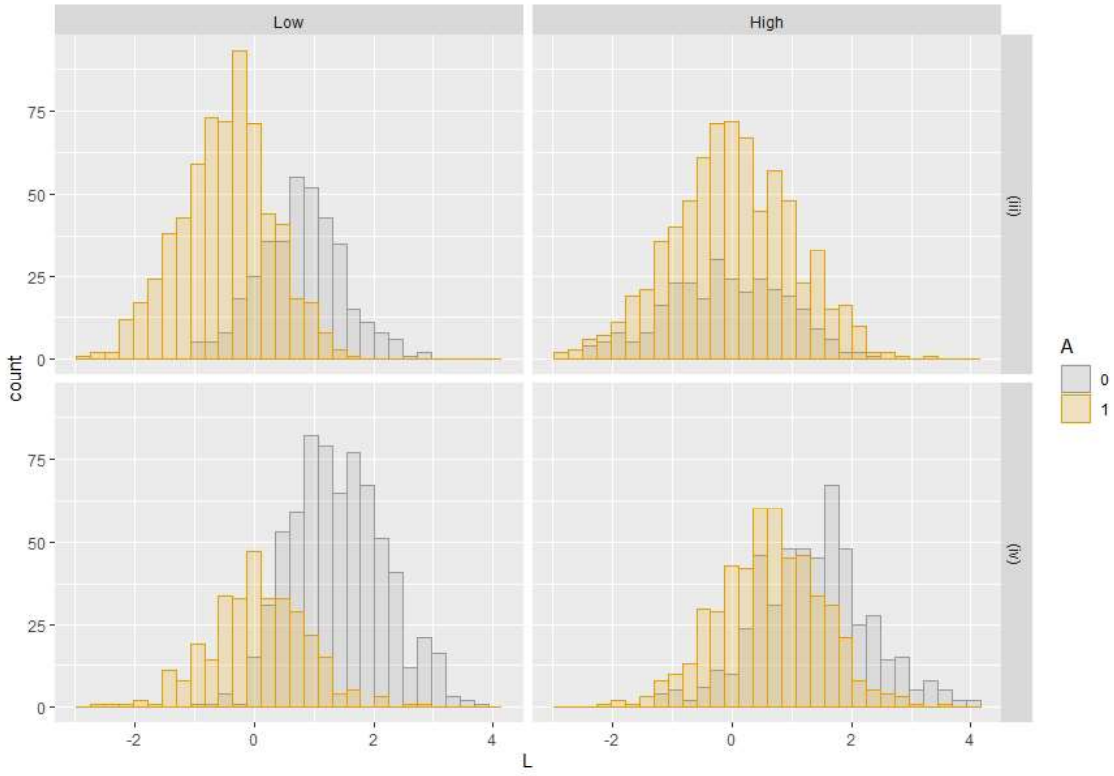
Covariate Overlap Simulation Studies

The main simulation study for scenarios (iii) and (iv) was repeated so that there was less overlap in the covariate distributions. In particular, to create low overlap, the parameter α_1 in the propensity score model was changed to -2.5 to decrease the amount of overlap in the distribution of L for $A = 0$ and $A = 1$. Web Figure 2 shows the empirical distribution of the covariate L by exposure A for data sets simulated under the original parameterization (denoted by "High" overlap) as well as the low overlap parameterization. The simulation study results in Web Table 6 show the SEE and naïve variance estimators are both biased and the corresponding Wald confidence intervals fail to cover at the nominal level when there is not sufficient covariate overlap. The bootstrap estimator described above in Web Appendix 3 yielded similar results, with $\widehat{ASE} = 0.12$ and coverage 0.79 for scenario (iii) Low and $\widehat{ASE} = 0.18$ and coverage 0.74 for scenario (iv) Low.

Web Table 6: Empirical standard error, average estimated standard error using the SEE and naïve variance estimates, 95% confidence interval coverage, and $\widehat{ASE}$ ratio (SEE/Naïve) for low and high overlap cases of each simulated scenario.

		SEE		Naïve		
Scenario	ESE	$\widehat{ASE}$	Coverage	$\widehat{ASE}$	Coverage	$\widehat{ASE}$ Ratio
(iii) Low	0.17	0.12	0.77	0.12	0.77	0.99
(iii) High	0.07	0.07	0.95	0.06	0.93	1.10
(iv) Low	0.30	0.18	0.75	0.22	0.84	0.83
(iv) High	0.11	0.10	0.94	0.15	1.00	0.65

$\widehat{ASE}$ = average estimated standard error; ESE = empirical standard error; SEE = stacked estimating equations.



Web Figure 2: Distribution of L by exposure A in low and high covariate overlap cases for Scenarios (iii) and (iv).

WEB APPENDIX REFERENCES

[1] D. B. Rubin, "Randomization analysis of experimental data: The fisher randomization test comment," *Journal of the American Statistical Association*, vol. 75, no. 371, pp. 591–593, 1980.

[2] R. Pirracchio, M. Carone, M. R. Rigon, E. Caruana, A. Mebazaa, and S. Chevret, "Propensity score estimators for the average treatment effect and the average treatment effect on the treated may yield very different estimates," *Statistical Methods in Medical Research*, vol. 25, no. 5, pp. 1938–1954, 2016.

[3] M. A. Hernán and J. M. Robins, "Estimating causal effects from epidemiological data," *Journal of Epidemiology & Community Health*, vol. 60, no. 7, pp. 578–586, 2006.

[4] T. Sato and Y. Matsuyama, "Marginal structural models as a tool for standardization," *Epidemiology*, vol. 14, no. 6, pp. 680–686, 2003.

[5] L. Stefanski and D. Boos, "The Calculus of M-Estimation," *The American Statistician*, vol. 56, no. 1, pp. 29–38, 2002.

[6] G. Casella and R. L. Berger, *Statistical Inference, Second Edition*. Duxbury Pacific Grove, CA, 2002.

[7] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

[8] M. P. Fay and B. I. Graubard, "Small-sample adjustments for wald-type tests using sandwich estimators," *Biometrics*, vol. 57, no. 4, pp. 1198–1206, 2001.

[9] S. Paul and X. Zhang, "Small sample gee estimation of regression parameters for longitudinal data," *Statistics in Medicine*, vol. 33, no. 22, pp. 3869–3881, 2014.

[10] P. Li and D. T. Redden, "Small sample performance of bias-corrected sandwich estimators for cluster-randomized trials with binary outcomes," *Statistics in Medicine*, vol. 34, no. 2, pp. 281–296, 2015.