Web supplement: Estimating and testing vaccine sieve effects using machine learning

1 Hazard-based TMLE estimator

Equation (5) in Appendix A relates the cause-specific hazard functions to the covariate-conditional cumulative incidence and suggests that an alternative plug-in estimator of cumulative incidence could be developed based on estimators of $\lambda := \{\lambda_{j,t} : t = 1, \ldots, t_0; j = 1, 2\}$. Such an estimator is a natural extension of Moore and van der Laan (2009) to the competing risks setting. To estimate the cause-specific hazard function for matched endpoints, we can use a pooled estimate as in Stitelman and van der Laan (2010). In particular, this estimate may be based on super learner.

To estimate the hazard function for mismatched endpoints, we require a minor modification to ensure that $\lambda_{1,t} + \lambda_{2,t} \leq 1$ for all t. In this case, using an estimate $\lambda_{n,1,t}$ of $\lambda_{1,t}$, we may regress pseudo-outcome $\tilde{M}_n(t) := M(t)/\{1 - \lambda_{n,1,t}(W)\}$ on W in the subset of observations with $Z = 1$ and $C(t-1) = N(t) = M(t-1) = 0$. This provides an estimate of $\tilde{\lambda}_{2,t}(w) := \text{pr}_{P_0}(T = t, J = 2 \mid Z = 1, C(t-1) = M(t-1) = N(t) = 0, W = w)$, which we can relate back to $\lambda_{2,t}(w)$ using the relationship $\lambda_{2,t}(w) = \tilde{\lambda}_{2,t}(w)/\{1 - \lambda_{1,t}(w)\}$.

The EIF can be written in terms of λ rather than μ. We define

$$Q_t(\lambda)(w) := \sum_{s=t+1}^{t_0} \left[\lambda_{1,s}(w) \prod_{m=t+1}^{s-1} \{1 - \lambda_{1,s}(w) - \lambda_{2,s}(w)\} \right],$$

and $R_t(o) := I(z = 1, c(t-1) = 0, n(t-1) = 0, m(t-1) = 0)$. The EIF can then be written as $D^*(\lambda, G, \zeta, \pi) = \sum_{j=1}^{2} \sum_{t=1}^{t} D_{t,j}(\lambda, \zeta, \pi) + D_W(\lambda, G)$, where for $t = 1, \ldots, t_0$ we set

$$\begin{aligned}
D_{t,1}(\lambda, \zeta, \pi)(o) &:= \frac{R_t(o)}{\zeta(w) \prod_{s=1}^{t-1} \pi_{0,s}(w)} \{1 - Q_t(\lambda)(w)\} \{n(t) - \lambda_{1,t}(w)\} \ , \\
D_{t,2}(\lambda, \zeta, \pi)(o) &:= -\frac{R_t(o)}{\zeta(w) \prod_{s=1}^{t-1} \pi_{0,s}(w)} Q_t(\lambda)(w) \{m(t) - \lambda_{2,t}(w)\} \ , \\
D_W(\lambda, G)(o) &:= Q_1(\lambda)(w) - \int Q_1(\lambda)(u) dG(u) \ .
\end{aligned}$$

A hazard-based TMLE can be constructed by choosing stopping criteria $c_n = o_P(n^{-1/2})$ and proceeding as follows:

1. If the conditional treatment probability ζ is known, set $\zeta_n = \zeta$; otherwise, construct estimate ζ_n of ζ. Construct estimate π_n of the censoring probabilities π.
2. Construct estimate $\lambda_n^0 = (\lambda_{n,1}^0, \lambda_{n,2}^0)$ of λ as described above.
3. Set $k = 0$. While $\left| \frac{1}{n} \sum_{i=1}^{n} D^*(\lambda_n^k, G_n, \zeta_n, \pi_n)(O_i) \right| > c_n$, repeat the following:

i) Fit logistic regression pooled over $t = 1, \ldots, t_0$ with outcome $N(t)$, offset $\text{logit}\{\lambda^k_{n,1,t}(W)\}$ and single covariate $\{\zeta_n(W)\prod_{s=1}^{t-1}\pi_{n,s}(W)\}^{-1}\{1 - Q_t(\lambda^k_n)(W)\}$ using only observations with $Z = 1$ and $C(t-1) = N(t-1) = M(t-1) = 0$. Denote by $\epsilon^k_{1,n}$ the MLE of the regression coefficient in this model. Define the updated estimate

$$\lambda^{k+1}_{n,1,t} := \text{expit}\left\{\text{logit}(\lambda^k_{n,1,t}) + \epsilon^k_{1,n}\frac{1 - Q_t(\lambda^k_n)}{\zeta_n\prod_{s=1}^{t-1}\pi_{n,s}}\right\}.$$

Define $\lambda^{k,\circ}_n := (\lambda^{k+1}_{1,n}, \lambda^k_{2,n})$.

ii) Fit logistic regression pooled over $t = 1, \ldots, t_0$ with outcome $M(t)/\{1-\lambda^{k+1}_{n,1,t}(W)\}$, offset term $\text{logit}[\lambda^k_{2,n,t}(W)/\{1 - \lambda^{k+1}_{n,1,t}(W)\}]$, and single covariate $-[\zeta_n(W)\prod_{s=1}^{t-1}\pi_{n,s}(W)\{1 - \lambda^{k+1}_{n,1,t}(W)\}]^{-1}Q_t(\lambda^{k,\circ}_n)(W)$ in the subset of data with $Z = 1, C(t-1) = 0, N(t-1) = 0, M(t-1) = 0$. Denote by $\epsilon^k_{2,n}$ the MLE of the regression coefficient in this model. Define the updated estimate

$$\lambda^{k+1}_{n,2,t} := (1 - \lambda^{k+1}_{n,1,t})\text{expit}\left\{\text{logit}\left(\frac{\lambda^k_{n,2,t}}{1 - \lambda^{k+1}_{n,1,t}}\right) - \epsilon^k_{2,n}\frac{Q_t(\lambda^{k,\circ}_n)}{\zeta_n\prod_{s=1}^{t-1}\pi_{n,s}(1 - \lambda^{k+1}_{n,1,t})}\right\}.$$

iii) Set $k = k + 1$.

4. Set $\lambda^*_n = \lambda^k_n$ and construct estimate by averaging over observed values of W,

$$\bar{\mu}^*_n = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{t_0}\left[\lambda^*_{n,1,t}(W_i)\prod_{s=1}^{t-1}\left\{1 - \lambda^*_{n,1,s}(W_i) - \lambda^*_{n,2,s}(W_i)\right\}\right].$$

In simulation studies, we have found the hazard-based TMLE to perform approximately as well as the iterated mean-based TMLE. However, van der Laan and Gruber (2012) argue that the mean-based approach may be preferable in practice since the iterated means are of reduced dimensionality relative to the collection of hazards, and thus may be easier to estimate well. We do note, however, that the mean-based TMLE may be more computationally intensive, particularly in data structures involving many time-points. In this case, the iterated mean TMLE requires performing computationally intensive regression (e.g., super learner) at each time-point, whereas the hazard-based TMLE requires only two such regressions that pool over all time-points. In such situations, the hazard-based approach may be preferable.

2 Asymptotic variance of vaccine efficacy and vaccine sieve effect estimators

Let $\boldsymbol{F}_n(t_0) := (F_{n,1,0}(t_0), F_{n,1,1}(t_0), F_{n,2,0}(t_0), F_{n,2,1}(t_0))^\top$ denote the TMLE estimator of $\boldsymbol{F}(t_0) := (F_{1,0}(t_0), F_{1,1}(t_0), F_{2,0}(t_0), F_{2,1}(t_0))^\top$. Let $\mu_{j,z} := \{\mu_{j,z,t_0} : t\}$ denote the true value for the iterated means used to identify $F_{j,z}(t_0)$ for $z = 0, 1$ and $j = 1, 2$, and let $\mu_{n,j,z}$ denote an estimator of these means. Similarly, let $\zeta_z(w) := \mathrm{pr}_{P_0}(Z = z \mid W = w)$ for $z = 0, 1$, and let $\pi_{z,t}(w) := \mathrm{pr}_{P_0}\{C(t) = 0 \mid Z = z, N(t-1) = 0, M(t-1) = 0, C(t-1) = 0, W = w\}$ for $z = 0, 1$ and $t = 1, \ldots, t_0 - 1$. Let $D^*_{j,z}(\eta_{j,z})$ denote the influence function of $F_{n,j,z}(t_0)$ from Theorem 2, where we introduce the shorthand $\eta_{j,z} := (\mu_{j,z}, \zeta_z, \pi_z, G)$. By asymptotic linearity of each component of $\boldsymbol{F}_n(t_0)$, we have that $\boldsymbol{F}_n(t_0)$ is itself asymptotically linear with influence function $\boldsymbol{D}^*(\eta) := (D^*_{1,0}(\eta_{1,0}), D^*_{1,1}(\eta_{1,1}), D^*_{2,0}(\eta_{2,0}), D^*_{2,1}(\eta_{2,1}))^\top$, where we defined $\eta := \{\eta_{j,z} : j, z\}$. By the multivariate central limit theorem, $n^{1/2}\{\boldsymbol{F}_n(t_0) - \boldsymbol{F}(t_0)\}$ converges in distribution to a four-dimensional multivariate normal variate with mean zero and covariance matrix $\Sigma := \mathrm{E}_{P_0}\{\boldsymbol{D}^*(\eta)(O)\boldsymbol{D}^*(\eta)(O)^\top\}$. The covariance matrix Σ is consistenly estimated by $\Sigma_n := \frac{1}{n}\sum_{i=1}^n \boldsymbol{D}^*(\eta_n)(O_i)\boldsymbol{D}^*(\eta_n)(O_i)^\top$, where $\eta_n := \{\eta_{n,j,z} : j, z\}$ is an estimate of η.

The asymptotic variance of the TMLE estimator of $VE_1(t_0)$ may be computed by noting that $VE_1(t_0) = h(\boldsymbol{F}(t_0))$. The gradient of h is $\nabla h(\boldsymbol{F}(t_0)) := (F_{1,1}(t_0)/F_{1,0}(t_0)^2, -1/F_{1,0}(t_0), 0, 0)^\top$. By the delta method, $n^{1/2}\{VE_{n,1}(t_0) - VE_1(t_0)\}$ converges in distribution to a mean-zero normally distributed variate with variance $\nabla h(\boldsymbol{F}(t_0))^\top \Sigma \nabla h(\boldsymbol{F}(t_0))$. An estimator of the asymptotic variance of $VE_{n,1}(t_0)$ is $\nabla h(\boldsymbol{F}_n)^\top \Sigma_n \nabla h(\boldsymbol{F}_n)$. A Wald-type interval could be constructed using this variance estimator, though we instead propose to invert a Wald-type interval constructed on the logarithmic scale. Specifically, we propose as $100 \times (1-\alpha)\%$ confidence interval for $VE_1(t_0)$ the interval

$$1 - \exp\left[\log\left\{\frac{F_{n,1,1}(t_0)}{F_{n,1,0}(t_0)} \pm z_{1-\alpha/2}\frac{\tau_n}{n^{1/2}}\right\}\right],$$

where τ_n^2 is an estimate of the asymptotic variance of $\log\{1 - VE_{n,1}(t_0)\}$, which can be computed using similar delta method arguments as above. Similar calculations can be used to produce a confidence interval for VE_2.

The asymptotic variance of the TMLE estimator of $\log VSE(t_0)$ may be computed by noting that $\log VSE(t_0) = f(\boldsymbol{F}(t_0))$, where the gradient of f is

$$\nabla f(\boldsymbol{F}(t_0)) := (F_{1,0}(t_0)^{-1}, -F_{1,1}(t_0)^{-1}, -F_{2,0}(t_0)^{-1}, F_{2,1}(t_0)^{-1})^\top .$$

The delta method implies that $n^{-1/2}\{\log VSE_n(t_0) - \log VSE(t_0)\}$ converges in distribution to a mean-zero normally distributed variate with variance $\nabla f(\boldsymbol{F}_0)^\top \Sigma \nabla f(\boldsymbol{F})$. Thus, we obtain a $100 \times (1-\alpha)\%$ confidence interval for $VSE(t_0)$ based upon $VSE_n(t_0)$ to be $\exp\{\log VSE_n(t_0) \pm z_{1-\alpha/2}\nu_n n^{-1/2}\}$ with $\nu_n^2 := \nabla f(\boldsymbol{F}_n)^\top \Sigma_n \nabla f(\boldsymbol{F}_n)$.

These results easily extend to the multiple outputation setting. Let $\boldsymbol{F}_{n,b}(t_0) := (F_{n,1,0,b}(t_0), F_{n,1,1,b}(t_0), F_{n,2,0,b}(t_0), F_{n,2,1,b}(t_0))^\top$ denote the vector of estimated cumulative incidences for the b-th outputed data set. Let $\boldsymbol{F}_{n,\mathrm{MO}}(t_0) := (\boldsymbol{F}_{n,1}(t_0), \ldots, \boldsymbol{F}_{n,B}(t_0))^\top$ denote a vector comprised of the B outputed cumulative incidence estimates. Let $\eta_{n,b}$ denote the estimates of η_0 based on the b-th outputed data set, and let $\boldsymbol{D}^*(\eta_{n,b})$ denote the vector of efficient influence functions evaluated at the estimates from the b-th outputed data set. Let $\boldsymbol{D}^*_{\mathrm{MO}}(\eta_n) := (\boldsymbol{D}^*(\eta_{n,1}), \ldots, \boldsymbol{D}^*(\eta_{n,B}))^\top$ denote the efficient influence function vector over all outputed data sets. Using nearly identical delta method calculus above, we can derive variance estimators for $\overline{VE}_{j,n}(t_0)$ and $\overline{VSE}_n(t_0)$.

References

Moore, K. and van der Laan, M. (2009). Increasing power in randomized trials with right censored outcomes through covariate adjustment. *Journal of Biopharmaceutical Statistics*, 19(6):1099–1131.

Stitelman, O. M. and van der Laan, M. J. (2010). Collaborative targeted maximum likelihood for time to event data. *The International Journal of Biostatistics*, 6(1).

van der Laan, M. J. and Gruber, S. (2012). Targeted minimum loss based estimation of causal effects of multiple time point interventions. *The International Journal of Biostatistics*, 8(1):1–34.