On Estimation of Vaccine Efficacy Using Validation Samples with Selection Bias: Supplementary Material

DANIEL O. SCHARFSTEIN[1], M. ELIZABETH HALLORAN[2],
HAITAO CHU[3], MICHAEL J. DANIELS[4]
[1] *Department of Biostatistics,*
Johns Hopkins Bloomberg School of Public Health, Baltimore MD 21205 USA
dscharf@jhsph.edu, phone: 410-955-2420, fax: 410-955-0958
[2] *Program in Biostatistics and Biomathematics, Fred Hutchinson Cancer Research Center*
Seattle, WA 98109 USA
[2] *Department of Biostatistics, University of Washington.*
Seattle, WA 98195 USA
[3] *Department of Epidemiology,*
Johns Hopkins Bloomberg School of Public Health, Baltimore MD 21205 USA
[4] *Department of Statistics,*
University of Florida, Gainesville, FL 32611 USA

February 28, 2006

A Large Sample Theory

Let

$$\begin{aligned} W_{z,x,y}(O) &= I(Z=z, X=x, A=1, R=1, Y=y) \\ \omega_{z,x,y} &= P[Z=z, X=x, A=1, R=1, Y=y] \\ V_{z,x,a}(O) &= I(Z=z, X=x, A=a) \\ \nu_{z,x,a}(O) &= P[Z=z, X=x, A=a] \end{aligned}$$

and

$$\begin{aligned} W(O) &= (W_{0,0,0}(O), W_{0,0,1}(O), W_{1,0,0}(O), W_{1,0,1}(O),\ W_{0,1,0}(O), W_{0,1,1}(O), W_{1,1,0}(O), W_{1,1,1}(O), \\ &\quad W_{0,2,0}(O), W_{0,2,1}(O), W_{1,2,0}(O), W_{1,2,1}(O))' \\ V(O) &= (V_{0,0,0}(O), V_{0,0,1}(O), V_{1,0,0}(O), V_{1,0,1}(O),\ V_{0,1,0}(O), V_{0,1,1}(O), V_{1,1,0}(O), V_{1,1,1}(O), \end{aligned}$$

$$
\begin{aligned}
& V_{0,2,0}(O), V_{0,2,1}(O), V_{1,2,0}(O), V_{1,2,1}(O))' \\
U(0) &= (W(0)', V(0)')' \\
\omega &= (\omega_{0,0,0}, \omega_{0,0,1}, \omega_{1,0,0}, \omega_{1,0,1}, \omega_{0,1,0}, \omega_{0,1,1}, \omega_{1,1,0}, \omega_{1,1,1}, \omega_{0,2,0}, \omega_{0,2,1}, \omega_{1,2,0}, \omega_{1,2,1})' \\
\nu &= (\nu_{0,0,0}, \nu_{0,0,1}, \nu_{1,0,0}, \nu_{1,0,1}, \nu_{0,1,0}, \nu_{0,1,1}, \nu_{1,1,0}, \nu_{1,1,1}, \nu_{0,2,0}, \nu_{0,2,1}, \nu_{1,2,0}, \nu_{1,2,1})' \\
\mu &= (\omega', \nu')'
\end{aligned}
$$

By the multivariate central limit theorem, we know that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{U(O_i) - \mu\} = \sqrt{n}(\tilde{\mu} - \mu) \xrightarrow{D} N_{24}(0, \Sigma)$$

where $\Sigma = E[(U(O) - \mu)(U(O) - \mu)']$ and $\tilde{\mu}$ has components

$$
\begin{aligned}
\tilde{\omega}_{z,x,y} &= \tilde{P}[Z = z, X = x, A = 1, R = 1, Y = y] \\
\tilde{\nu}_{z,x,a} &= \tilde{P}[Z = z, X = x, A = a]
\end{aligned}
$$

The asymptotic variance can be estimated $\hat{\Sigma} = \tilde{E}[(U(O_i) - \tilde{\mu})(U(O_i) - \tilde{\mu})']$, where $\tilde{E}[\cdot]$ is the empirical expectation operator.

For fixed $\beta_{z,x}$'s, we see that

$$
\begin{aligned}
VE_{S,x} &= 1 - \exp(f_x(\mu)) \\
\hat{VE}_{S,x} &= 1 - \exp(f_x(\tilde{\mu})) \\
VE_S &= 1 - \exp(g(\mu)) \\
\hat{VE}_S &= 1 - \exp(g(\tilde{\mu}))
\end{aligned}
$$

where

$$
\begin{aligned}
f_x(\mu) &= \log(\omega_{1,x,1}) + \log(\nu_{1,x,1}) - \log(\nu_{1,x,1} + \nu_{1,x,0}) - \log(\beta_{1,x}\omega_{1,x,0} + \omega_{1,x,1}) - \\
&\quad \log(\omega_{0,x,1}) - \log(\nu_{0,x,1}) + \log(\nu_{0,x,1} + \nu_{0,x,0}) + \log(\beta_{0,x}\omega_{0,x,0} + \omega_{0,x,1}) \\
g(\mu) &= \log\left(\sum_{x=0}^{2}\left\{\frac{\omega_{1,x,1}\frac{\nu_{1,x,1}}{\nu_{1,x,1}+\nu_{1,x,0}}}{\beta_{1,x}\omega_{1,x,0} + \omega_{1,x,1}}\left(\sum_{z=0}^{1}\sum_{a=0}^{1}\nu_{z,x,a}\right)\right\}\right) - \\
&\quad \log\left(\sum_{x=0}^{2}\left\{\frac{\omega_{0,x,1}\frac{\nu_{0,x,1}}{\nu_{0,x,1}+\nu_{0,x,0}}}{\beta_{0,x}\omega_{0,x,0} + \omega_{0,x,1}}\left(\sum_{z=0}^{1}\sum_{a=0}^{1}\nu_{z,x,a}\right)\right\}\right)
\end{aligned}
$$

By the multivariate delta method, we can now derive the asymptotic distribution of $f_x(\tilde{\mu})$ and $g(\tilde{\mu})$, which we then use to construct confidence intervals for $VE_{S,x}$ and VE_S using the $1-\exp(\cdot)$ function. In particular, we know that $f_x(\tilde{\mu}) - f_x(\mu) \approx N(0, \hat{\sigma}^2_{f_x})$ and $g(\tilde{\mu}) - g(\mu) \approx N(0, \hat{\sigma}^2_g)$,

where

$$\begin{aligned}\hat{\sigma}^2_{f_x} &= \frac{1}{n} \bigtriangledown f_x(\tilde{\mu})' \hat{\Sigma} \bigtriangledown f_x(\tilde{\mu}) \\ \hat{\sigma}^2_{g} &= \frac{1}{n} \bigtriangledown g(\tilde{\mu})' \hat{\Sigma} \bigtriangledown g(\tilde{\mu})\end{aligned}$$

and $\bigtriangledown f_x(\mu)$ and $\bigtriangledown g(\mu)$ are the partial derivative matrices (with respect to μ) of the functions $f_x(\cdot)$ and $g(\cdot)$, respectively.

So, 95% confidence intervals for $VE_{S,x}$ and V_S are given by

$$\begin{aligned}&\left[1 - \exp\left(f_x(\tilde{\mu}) + 1.96\hat{\sigma}_{f_x}\right), 1 - \exp\left(f_x(\tilde{\mu}) - 1.96\hat{\sigma}_{f_x}\right)\right] \\ &\left[1 - \exp\left(g(\tilde{\mu}) + 1.96\hat{\sigma}_{g}\right), 1 - \exp\left(g(\tilde{\mu}) - 1.96\hat{\sigma}_{g}\right)\right],\end{aligned}$$

respectively.

B Sampling from the posterior distribution

We consider the complete data as $\boldsymbol{\mathcal{F}} = \{F_i = (Z_i, X_i, A_i, R_i, Y_i : A_i = 1) : i = 1, \ldots, n\}$. In the full data, there is no missing data on true influenza for those who have MAARI. Let $\boldsymbol{\mathcal{F}} - \boldsymbol{\mathcal{O}}$ denote the missing data. In the sampling from the posterior distribution, we use data augmentation (Tanner and Wong, 1987). Specifically, we will augment the posterior by $\mathcal{F} - \mathcal{O}$ and a vector of positive latent variables $\boldsymbol{u}$ with independent components $u_{z,x}$ (Damien *et al.* 1999). We will seek to compute the posterior,

$$\pi(\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{p}, \boldsymbol{\phi}, \boldsymbol{\mathcal{F}} - \boldsymbol{\mathcal{O}} | \boldsymbol{\mathcal{O}}) = \pi(\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{p}, \boldsymbol{\phi} | \boldsymbol{\mathcal{F}}) \pi(\boldsymbol{\mathcal{F}} - \boldsymbol{\mathcal{O}} | \boldsymbol{\mathcal{O}}),$$

where

$$\pi(\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{p}, \boldsymbol{\phi} | \boldsymbol{\mathcal{F}}) \propto \mathcal{L}(\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{\eta}; \boldsymbol{\mathcal{F}}) \mathcal{L}(\boldsymbol{p}; \boldsymbol{\mathcal{F}}) \mathcal{L}(\boldsymbol{\phi}; \boldsymbol{\mathcal{F}}) \pi(\boldsymbol{\beta}) \pi(\boldsymbol{\eta} | \boldsymbol{\beta}) \pi(\boldsymbol{p}) \pi(\boldsymbol{\phi})$$

$$\begin{aligned}u^*_{z,x}(\beta_{z,x}, \eta_{z,x}; \boldsymbol{\mathcal{F}}) &= \left\{ \{\beta_{z,x}\eta_{z,z}\}^{\sum_{i \in \mathcal{S}_{z,x}} A_i R_i Y_i} \{1 - \beta_{z,x}\eta_{z,x}\}^{\sum_{i \in \mathcal{S}_{z,x}} A_i (1-R_i) Y_i} I(0 \leq \beta_{z,x}\eta_{z,x} \leq 1) \right\} \\ \mathcal{L}(\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{\eta}; \boldsymbol{\mathcal{F}}) &= \prod_{z=0}^{1} \prod_{x=0}^{2} \left\{ I(u_{z,x} \leq u^*_{z,x}(\beta_{z,x}, \eta_{z,x}; \boldsymbol{\mathcal{F}})) \{\eta_{z,x}\}^{\sum_{i \in \mathcal{S}_{z,x}} A_i R_i (1-Y_i)} \{1 - \eta_{z,x}\}^{\sum_{i \in \mathcal{S}_{z,x}} A_i (1-R_i)(1-Y_i)} \right. \\ &\qquad \left. I\left(0 \leq \eta_{z,x} \leq \min\left\{1/\beta_{z,x}, 1\right\}\right) \right\} \qquad (1) \\ \mathcal{L}(\boldsymbol{p}; \boldsymbol{\mathcal{F}}) &= \prod_{z=0}^{1} \prod_{x=0}^{2} \left\{ \{p_{z,x}\}^{\sum_{i \in \mathcal{S}_{z,x}} A_i Y_i} \{1 - p_{z,x}\}^{\sum_{i \in \mathcal{S}_{z,x}} A_i (1-Y_i)} I(0 \leq p_{z,x} \leq 1) \right\} \\ \mathcal{L}(\boldsymbol{\phi}; \boldsymbol{\mathcal{F}}) &= \prod_{z=0}^{1} \left\{ \left\{ \prod_{x=0}^{2} \{\phi_{z,x,1}\}^{\sum_{i \in \mathcal{S}_{z,x}} A_i} \{\phi_{z,x,0}\}^{\sum_{i \in \mathcal{S}_{z,x}} (1-A_i)} I(0 \leq \phi_{z,x,1}, \phi_{z,x,0} \leq 1) \right\} \right. \\ &\qquad \left. I\left(\sum_{x=0}^{2} \sum_{a=0}^{1} \phi_{z,x,a} = 1 \right) \right\}\end{aligned}$$

and $\mathcal{S}_{z,x} = \{i : Z_i = z, X_i = x\}$. We will obtain the posterior $\pi(\boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{p}, \boldsymbol{\phi}|\boldsymbol{\mathcal{O}})$ by integrating out $\boldsymbol{\mathcal{F}} - \boldsymbol{\mathcal{O}}$ and $\boldsymbol{u}$ (via Monte Carlo integration within the sampling algorithm).

To sample from $\pi(\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{p}, \boldsymbol{\phi}, \boldsymbol{\mathcal{F}} - \boldsymbol{\mathcal{O}}|\boldsymbol{\mathcal{O}})$, we sequentially sample from all of the full conditionals as follows:

1. $\pi(\boldsymbol{p}|\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\phi}, \boldsymbol{\mathcal{F}}) = \pi(\boldsymbol{p}|\boldsymbol{\mathcal{F}})$

   The distribution of $\boldsymbol{p}$ given $\boldsymbol{\mathcal{F}}$ can be shown to have indepdendent components where

   $$p_{z,x}|\boldsymbol{\mathcal{F}} \sim \text{Beta}\left(\sum_{i \in S_{z,x}} A_i Y_i + 1, \sum_{i \in S_{z,x}} A_i(1 - Y_i) + 1\right)$$

2. $\pi(\boldsymbol{\mathcal{F}} - \boldsymbol{\mathcal{O}}|\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{p}, \boldsymbol{\phi}, \boldsymbol{\mathcal{O}})$

   For the set of individuals in stratum $S_{z,x}$ with $R_i = 0$ and $A_i = 1$, we draw $S^{\dagger}_{z,x} = \{Y_i : R_i = 0, A_i = 1, Z_i = z, X_i = x\}$ from a binomial distribution with success probability

   $$\frac{p_{z,x}}{p_{z,x} + (1 - p_{z,x})(1 - \eta_{z,x})}$$

   truncated from above by the condition $u_{z,x} < u^*_{z,x}(\beta_{z,x}, \eta_{z,x}; (\boldsymbol{\mathcal{O}}, S^{\dagger}_{z,x}))$.

3. $\pi(\boldsymbol{u}|\boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{p}, \boldsymbol{\phi}, \boldsymbol{\mathcal{F}}) = \prod_{z=0}^{1} \prod_{x=0}^{2} \pi(u_{z,x}|\boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\mathcal{F}})$

   For each z, x stratum, $\pi(u_{z,x}|\boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\mathcal{F}})$ is uniform on the interval $(0, u^*_{z,x}(\beta_{z,x}, \eta_{z,x}, \boldsymbol{\mathcal{F}}))$.

4. $\pi(\boldsymbol{\eta}, \boldsymbol{\beta}|\boldsymbol{u}, \boldsymbol{p}, \boldsymbol{\phi}, \boldsymbol{\mathcal{F}}) \equiv \pi(\boldsymbol{\eta}, \boldsymbol{\beta}|\boldsymbol{u}, \boldsymbol{\mathcal{F}})$

   We will use slice sampling (Neal, 2003) to sample from the full conditional distribution of $(\boldsymbol{\beta}, \boldsymbol{\eta})$. Given the form of (1), we will sample from a truncated form of $\pi(\boldsymbol{\beta})$ and a truncated Beta$(\sum_{i \in S_{z,x}} A_i R_i (1 - Y_i) + 1, \sum_{i \in S_{z,x}} A_i (1 - R_i)(1 - Y_i) + 1)$ for $\eta_{z,x}$. The truncation bounds cannot be entirely expressed in closed form. We now describe how to obtain them sequentially.

   We want to sample the $(\beta_{z,x}, \eta_{z,x})$ in pairs, but $\pi(\boldsymbol{\beta}) \neq \prod_{z=0}^{1} \prod_{x=0}^{2} \pi(\beta_{z,x})$. To deal with this, we factor $\pi(\boldsymbol{\beta})$ as $\prod_{z=0}^{1} \prod_{x=0}^{2} \pi^{\star}(\beta_{z,x})$ where

   $$\pi^{\star}(\beta_{z,x}) = \pi(\beta_{z,x}|\{\beta_{z^{\star},x^{\star}} : z^{\star} = z, x^{\star} < x \text{ and } z^{\star} < z\})$$

   and $\pi^{\star}(\beta_{0,0}) = \pi(\beta_{0,0})$. Then, in that order, we sample the pairs $(\beta_{z,x}, \eta_{z,x})$. To sample $\beta_{z,x}$, we note that $\pi^{\star}(\beta_{z,x})$ is truncated by the condition $I(u_{z,x} < u^*_{z,x}(\gamma_{z,x}; \boldsymbol{\mathcal{F}}))$, where $u^*_{z,x}(\gamma_{z,x}; \boldsymbol{\mathcal{F}}) = u^*(\beta_{z,x}, \eta_{z,x}; \boldsymbol{\mathcal{F}})$ with $\gamma_{z,x} = \beta_{z,x}\eta_{z,x}$. Let

   $$\hat{\gamma}_{z,x} = \frac{\sum_{i \in \mathcal{S}_{z,x}} A_i R_i Y_i}{\sum_{i \in \mathcal{S}_{z,x}} A_i Y_i}$$

denote the maximum likelihood estimate of $u^*(\gamma_{z,x}; \boldsymbol{\mathcal{F}})$. Note that $u_{z,x} \leq u^*(\hat{\gamma}_{z,x}; \boldsymbol{\mathcal{F}})$.

Since we sample $\beta_{z,x}$ before $\eta_{z,x}$, we need to make sure we sample $\beta_{z,x}$ such that there are values of $\eta_{z,x}$ which will satisfy this condition. If $0 < \hat{\gamma}_{z,x} \leq 1$, then $\beta_{z,x}$ can lie in the interval $(\beta^*_{z,x}, \infty)$, where $\beta^*_{z,x}$ is equal to the unique $\gamma^L_{z,x}$ satisfying $u^*(\gamma^L_{z,x}; \boldsymbol{\mathcal{F}}) = u_{z,x}$ and $\gamma^L_{z,x} \leq \hat{\gamma}_{z,x}$. Since $u^*(\gamma_{z,x}; \boldsymbol{\mathcal{F}})$ is monotone in $\gamma_{z,x}$ for $0 < \gamma_{z,x} \leq \hat{\gamma}_{z,x}$, $\beta^*_{z,x}$ can be found by performing a bi-section search for the equality condition on the interval $(0, \hat{\gamma}_{z,x}]$. If $\hat{\gamma}_{z,x} = 0$, then $\beta_{z,x}$ can lie in $(0, \infty)$.

For $\eta_{z,x}$ given $\beta_{z,x}$, we sample from a truncated Beta distribution with parameters as specified above. For $0 \leq \hat{\gamma}_{z,x} < 1$, define $\gamma^U_{z,x}$ to be the unique quantity satisfying $u^*(\gamma^U_{z,x} : \boldsymbol{\mathcal{F}}) = u_{z,x}$ and $\gamma^U_{z,x} > \hat{\gamma}_{z,x}$. If $0 < \hat{\gamma}_{z,x} < 1$, $\eta_{z,x}$ must lie in the interval $\left[\frac{\gamma^L_{z,x}}{\beta_{z,x}}, \min(\frac{\gamma^U_{z,x}}{\beta_{z,x}}, 1)\right]$. If $\hat{\gamma}_{z,x} = 0$, $\eta_{z,x}$ must lie in the interval $\left[0, \min(\frac{\gamma^U_{z,x}}{\beta_{z,x}}, 1)\right]$. If $\hat{\gamma}_{z,x} = 1$, $\eta_{z,x}$ must lie in the interval $\left[\min(\frac{\gamma^L_{z,x}}{\beta_{z,x}}, 1), 1\right]$.

5. $\pi(\boldsymbol{\phi}|\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{\beta}, \boldsymbol{p}, \boldsymbol{\mathcal{F}}) = \prod_{z=0}^{1} \pi(\boldsymbol{\phi}_z|\mathcal{F})$

   Given $\boldsymbol{\mathcal{F}}$, sample $\boldsymbol{\phi}_z$ from a Dirichlet distribution with parameters $((1+n_{z1}, 1+n_{z2}, \ldots, 1+n_{zK}))$.

6. Compute $VE_{S,x}$ and VE_S using Equations (11) and (12) from the main manuscript, respectively.

Repeat steps 1-6 K times, discarding an appropriate number of samples from the burn-in period. Obtain the posterior mean, median, mode, and $1-\alpha$ HPD credible set or $1-\alpha$ equal tail credible set of $VE_{S,x}$ and VE_S from the Monte Carlo samples.

To perform Bayesian analysis with fix $\boldsymbol{\beta}$, we sample $\eta_{z,x}$ as described in Step 4. given the fixed value of $\boldsymbol{\beta}$.