# Web-based Supplementary Materials for A Bayesian Approach to Improved Estimation of Causal Effect Predictiveness for a Principal Surrogate Endpoint

by Corwin M. Zigler and Thomas R. Belin

## Appendix A: Details of the Data Generation for the Simulated Vaccine Trial with a Constant Biomarker

We follow the simulation scheme of GH that is designed to mimic the first preventive HIV vaccine efficacy trial. The candidate surrogate in this trial was 50% neutralization titers against the HIV recombinant gp120 molecule measured at $t = 1.5$ months post-baseline, and the primary outcome was HIV infection at 36 months follow-up. The constant biomarker in this case is the lower limit of detection of the antibody assay, $c = 0$, so $S(0) = 0$ for all patients. Simulated datasets contain 1,805 placebo recipients and 3,598 vaccine recipients, and we assume that immune biomarker data are subject to limit-of-detection left censoring. $(X, S(1))$ are simulated from a bivariate normal distribution with each component having a mean of 0.41 and a standard deviation of 0.55. We set the correlation between $X$ and $S(1)$, $\rho$, to be $0.5, 0.7$, or $0.9$ to represent scenarios where the pretreatment covariate predicts immune responses to varying degrees. Values of $S(1) < 0$ were truncated to 0 to reflect the limit-of-detection left censoring. The case-cohort sampling scheme follows from obtaining measured samples on all infected patients (cases) and from a subcohort of uninfected patents (controls) in both treatment arms. The ratio of controls:cases was 3:1 in both arms. Let $\delta_i$ be the indicator of whether patient $i$ is sampled in the case-cohort scheme. Values of $(X, S(1))$ were retained for all vaccine recipients with $\delta_i = 1$, and values $X$ were retained for all placebo recipients with $\delta_i = 1$.

Infection outcomes are simulated from $r_z(s_1, s_0, x) = P(Y(z) = 1|S(0) = s_0, S(1) = s_1, X = x) = \Phi(\beta_{z0} + \beta_{z1}s_1 + \beta_{z2}x)$. For the vaccine arm, we set $(\beta_{10}, \beta_{11}, \beta_{12})$ = (-1.21, -0.67, -0.1). For the placebo arm, we consider two scenarios reflecting an overall vaccine effect of 50% reduction in the number of infections: one where the surrogate has no value, corresponding to $\beta_{00} = -0.825, \beta_{01} = -0.67$, and $\beta_{02} = -0.1$ (scenario (a)), and one where it has high value, corresponding to $\beta_{00} = -1.1, \beta_{01} = 0.0$, and $\beta_{02} = -0.1$ (scenario (b)).

## Appendix B: Computational Details

With and without a constant biomarker, the complete-data likelihood is expressed as:

$$L(\theta) = \prod_{i=1}^{n} f(X_i, S_i(0), S_i(1)|\theta)\times$$
$$\prod_{i:Z_i=0}\{r_0(S_i(1), S_i(0), X_i)\}^{Y_i(0)}\{1-r_0(S_i(1), S_i(0), X_i)\}^{1-Y_i(0)}\times \tag{1}$$
$$\prod_{i:Z_i=1}\{r_1(S_i(1), S_i(0), X_i)\}^{Y_i(1)}\{1-r_1(S_i(1), S_i(0), X_i)\}^{1-Y_i(1)}$$

where the $r_z(s_1, s_0, x)$ are modeled with probit regressions. To ease computation, we represent these probit models with latent continuous variables $Y_i^*(Z_i)$,

$$Y_i(Z_i) = 1 \quad \text{if} \quad Y_i^*(Z_i) = g_z(S_i(1), S_i(0), X_i; \beta_z) + U_i(Z_i) > 0$$
$$Y_i(Z_i) = 0, \quad \text{otherwise}$$

where $U_i(Z_i) \sim N(0, 1)$.

The first term in (1) is modeled with a normal distribution, and the latent $Y_i^*$ allows specification of a single normal distribution for each other lines of (1). In the steps described below, all sampling is conditional on current parameter values and current sampled values of missing data.

### *B.1 MCMC details with a constant biomarker and case-cohort sampling*

We accommodate the limit-of-detection left censoring as described in Chib (1992). $N_+^0$ and $N_-^0$ denote normal distributions truncated to be above and below 0, respectively. For each analysis, three MCMC chains were run for 4,000 iterations, saving every 10th sample and discarding the first 2,000 as burn-in iterations. Convergence was checked visually and using the potential scale-reduction statistics ($\hat{R}$) (Gelman and Rubin, 1992); no presented analysis provided evidence against convergence, with a maximum value of $\hat{R}$ of 1.11 across all parameters of all simulations. The steps of the MCMC are outlined as follows:

(1) Sample $a_0, a_1, \sigma^2_{S(1)|X}$ using a standard Bayesian regression of $S_i(1)$ on $X_i$.

(2) Sample $\mu_X$ and $\frac{1}{\sigma_X^2}$ from normal and gamma distributions, respectively, using standard analysis tools.

(3) For $Z_i = z$ and $\delta_i = 0$, sample $X_i \sim N(m_{xzi}, v_{xz})$ where

$$m_{xzi} = \tfrac{\mu_X}{\sigma_X^2} + \tfrac{a_1(S_i(1)-a_0)}{\sigma^2_{S(1)|X}} + \beta_{z2}(Y_i^*(z) - \beta_{z0} - \beta_{z1}S_i(1)) \times v_{xz}$$

$$v_{xz} = (\tfrac{1}{\sigma_X^2} + \tfrac{a_1^2}{\sigma^2_{S(1)|X}} + \beta_{z2}^2)^{-1}$$

(4) For $Z_i = 1, \delta_i = 1$, and $S_i(1) = 0$ ( the limit of detection), sample $S_i(1) \sim N_{-}^{0}(a_0 + a_1X_i, \sigma^2_{S(1)|X})$.

(5) For patients with $Z_i = z$ and $z\delta_i = 0$, sample $S_i(1) \sim N(m_{zi}, v_z)$ where

$$m_{zi} = (\beta_{z1}(Y_i^*(z) - \beta_{z0} - \beta_{z2}X_i) + \tfrac{a_0+a_1X_i}{\sigma^2_{S(1)|X}}) \times v_z$$

$$v_z = (\tfrac{1}{\sigma^2_{S(1)|X}} + \beta_{z1}^2)^{-1}$$

(6) For all patients, sample the $Y_i^*(z)$ for $z = 0, 1$:

For $Y_i(z) = 0, Y_i^*(z) \sim N_{-}^{0}(\beta_{z0} + \beta_{z1}S_i(1) + \beta_{z2}X_i, 1)$

For $Y_i(z) = 1, Y_i^*(z) \sim N_{+}^{0}(\beta_{z0} + \beta_{z1}S_i(1) + \beta_{z2}X_i, 1)$

(7) Sample the $\beta$ parameters using a standard Bayesian regression:

$$Y_i(Z_i) = \beta_{10} + \beta_{11}S_i(1) + \beta_{12}X_i +$$

$$(\beta_{00} - \beta_{10})(1 - Z_i) + (\beta_{01} - \beta_{11})(1 - Z_i)S_i(1) + (\beta_{02} - \beta_{12})(1 - Z_i)X_i + U_i(Z_i)$$

(8) Calculate $EDE$, $EAE$, and $PAE$.

### *B.2 MCMC details with varying control-group response*

Sampling $\Sigma$ while holding fixed the elements corresponding to $\phi$ precludes the use of Wishart or Inverse-Wishart distributions. We sample each element of $\Sigma$ separately via a normal random-walk proposal distribution and a Metropolis step (Gelman, Carlin, Stern, and Rubin, 2004) subject to the constraint that $\Sigma$ remain positive definite. The normal proposal distributions are centered at the sampler's current parameter values with proposal variances that are obtained adaptively during a "pre burn-in" period of the sampler to ensure that the sampler moves efficiently through the parameter space (Roberts and Rosenthal, 2009).

Three MCMC chains were run and proposal variances adapted until Metropolis steps achieved approximately 44% acceptance, followed by 20,000 additional burn-in iterations and 20,000 more samples from which every 10th was used for inference. No analysis provided evidence against convergence, with $\hat{R} < 1.05$ for all parameters under all assumed values of $\phi$. The steps of the MCMC are as follows:

(1) Update $\mu = (\mu_{S(0)}, \mu_{S(1)}, \mu_X)$ from a multivariate normal distribution using standard analysis tools.

(2) For $j = 1, 2, 3$ and $k = 1, 2, 3$, sample a proposal for the $(j, k)$ element of $\Sigma$ from a normal distribution to generate $\Sigma^*$. Reject if the proposal value is $\leqslant 0$ or if $\Sigma^*$ is not positive definite, otherwise accept with probability:

$$exp[-\frac{4+n}{2}log(|\Sigma^*|) - \frac{1}{2}(tr(\Sigma^{*-1}SS_w)) + \frac{4+n}{2}log(|\Sigma|) + \frac{1}{2}(tr(\Sigma^{-1}SS_w))]$$

where $SS_w$ is the sum over $i$ of $W_iW_i'$ with $W_i = (S_i(0) - \mu_{S(0)}, S_i(1) - \mu_{S(1)}, X_i - \mu_X)'$ and $(k, j) \neq (1, 2)$ or $(2, 1)$.

(3) Update the $(1, 2)$ and $(2, 1)$ elements of $\Sigma$ with $\phi\sqrt{\sigma^2_{S(0)}\sigma^2_{S(1)}}$ where $\sigma^2_{S(z)}$ is the $(z+1, z+1)$ element of $\Sigma$, $z = 0, 1$, provided the resulting $\Sigma$ is positive definite.

(4) Sample the $S_i^{mis}$

For $Z_i = 0$, sample from $S_i(1) \sim N(m_{0i}v_0, v_0)$ where $v_0 = (s_{22} + \beta_{02}^2)^{-1}$ and

$$m_{0i} = \beta_{02}(Y_i^*(0) - \beta_{00} - (\beta_{01} - \beta_{02})(S_i(0) - \bar{S}(0)) - \beta_{03}(X_i - \bar{X}))+$$
$$\mu_{S(1)}s_{22} - (S_i(0) - \mu_{S(0)})s_{12} - (X_i - \mu_X)s_{23} + \beta_{02}^2\bar{S}(1)$$

For $Z_i = 1$, sample from $S_i(0) \sim N(m_{1i}v_1, v_1)$ where $v_1 = (s_{11} + (\beta_{12} - \beta_{11})^2)^{-1}$ and

$$m_{1i} = (\beta_{11} - \beta_{12})(Y_i^*(1) - \beta_{10} - \beta_{12}(S_i(1) - \bar{S}(1)) - \beta_{13}(X_i - \bar{X}))+$$
$$\mu_{S(0)}s_{11} - (S_i(1) - \mu_{S(1)})s_{12} - (X_i - \mu_X)s_{13} + (\beta_{11} - \beta_{12})^2\bar{S}(0)$$

Here, $s_{jk}$ denotes the $(j, k)$ element of $\Sigma^{-1}$ .

(5) Adapt steps (6) – (8) from Appendix B.1 accordingly.

## C Prior vs. Posterior distributions

### *C.1 Simulated scenario under the constant biomarker special case*

Figure 1 presents prior densities and posterior histograms of the $\beta$ parameters from the analysis of one dataset simulated with high surrogate value and $\rho = 0.5$. The prior specification dominates the posterior distribution of $(\beta_{02} - \beta_{12})$ reflecting prior belief in the absence of an interaction between $Z$ and $X$, although this interaction is not forced to be exactly 0 as in previously-used strategies. Prior vs. posterior plots from other simulated scenarios are not pictured, but appear similar to Figure 1 except for the increased flatness of the prior distribution for $(\beta_{01} - \beta_{11})$ in scenarios with higher $\rho$.

[Figure 1 about here.]

### *C.2 Data analysis of ACTG 320*

Figure 2 depicts prior densities and posterior histograms of the $\beta$ parameters for the analysis with $\phi = 0.4$. Examination of prior vs. posterior densities for other assumed values of $\phi$ (not pictured) produced similar results. Here we see that prior specification dominates the posterior distributions for $\beta_{03}$ and $\beta_{13} - \beta_{03}$, reflecting the the strong prior belief that the relationship between $X$ and $Y(z)$ estimated from (4) in the main text is correct.

[Figure 2 about here.]

## D Basic Graphical Checks of Multivariate Normality Assumption in the ACTG 320 Data Analysis

As basic summaries, we plot histograms of $X$ in all patients (Figure 3(a)), of $S(0)$ observed in $Z = 0$ patients (Figure 3(b)), and of $S(1)$ observed in $Z = 1$ patients (Figure 3(c)). We also provide scatterplots of $X$ vs. $S(0)$ in $Z = 0$ patients (Figure 4(a)) and of $X$ vs. $S(1)$ in $Z = 1$ patients (Figure 4(b))

[Figure 3 about here.]

[Figure 4 about here.]

## References

Chib, S. (1992). Bayes inference in the tobit censored regression model. *Journal of Econometrics* **51,** 79–99.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC, New York, 2nd edition.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **7,** 457–472.

Roberts, G. O. and Rosenthal, J. S. (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics* **18,** 349–367.

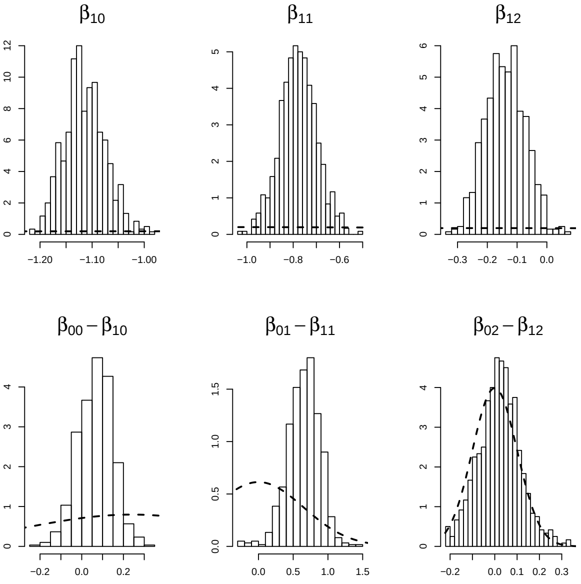


**Figure 1**: Prior densities (dashed) vs. posterior histograms for $\beta$ parameters from $r_z(s_1, c, x)$ from one simulation with a constant biomarker having high surrogate value and $\rho = 0.5$.

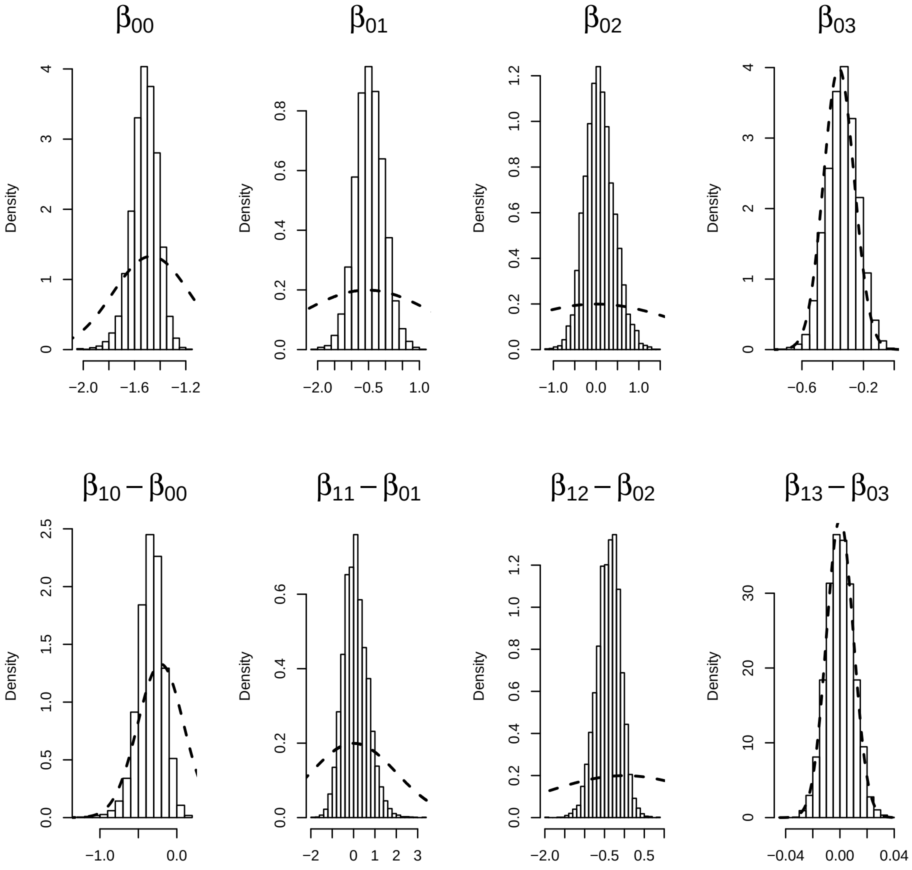
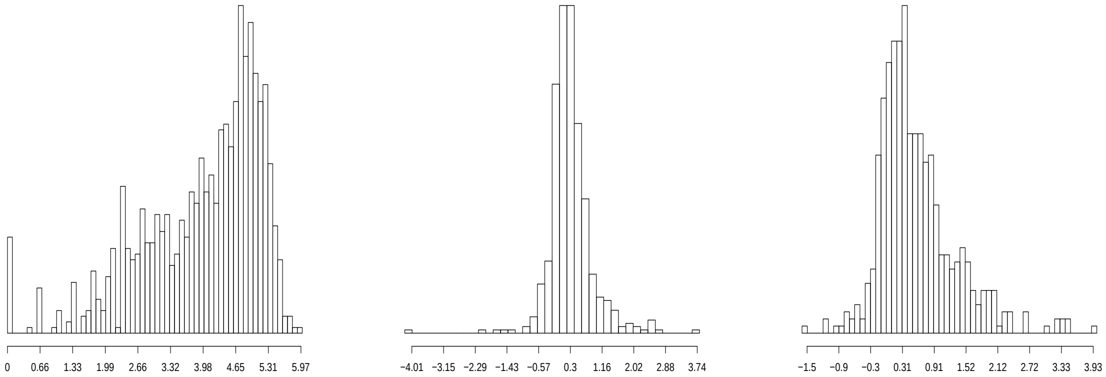


**Figure 2**: Prior densities (dashed) vs. posterior histograms for $\beta$ parameters from ACTG 320, $\phi = 0.4$.

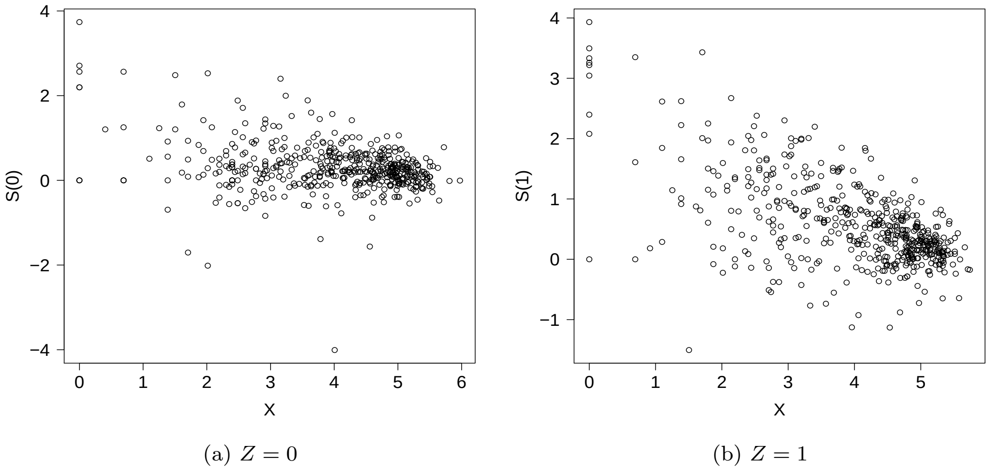


(a) $X$, baseline log(CD4)

(b) $S(0)$, change in log(CD4), $Z = 0$

(c) $S(1)$, change in log(CD4), $Z = 1$

**Figure 3**: Observed distributions of $X$, $S(0)$, and $S(1)$.



(a) $Z = 0$ (b) $Z = 1$

**Figure 4**: Bivariate distributions of $S(Z), X$ for $Z = 0, 1$.