Online supplementary material for

"Inverse Probability Weighted Cox Model in Multi-Site Studies without Sharing Patient-Level Data"

Di Shu[1], Kazuki Yoshida[2], Bruce H. Fireman[3], Sengwee Toh[1]

[1]Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, Massachusetts, USA

[2]Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA

[3]Division of Research, Kaiser Permanente Northern California, Oakland, California, USA

* Correspondence to Di Shu, Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, 401 Park Drive, Suite 401 East, Boston, MA 02215, USA (e-mail: Di_Shu@harvardpilgrim.org).

Appendix 1

In this proof we show that the robust sandwich variance estimate, given by equation (2) in the main manuscript, can be calculated using only summary-level quantities shared by the participating data-contributing sites.

We expand $\phi_i(k,\hat{\theta})$ and obtain

$$q = \sum_{k=1}^{K} \sum_{i \in \Omega_k} \phi_i^2(\mathrm{k},\hat{\theta})$$

$$= \sum_{k=1}^{K} q_1(k) + \sum_{k=1}^{K} q_2(k) + \sum_{k=1}^{K} q_3(k) - 2\sum_{k=1}^{K} q_4(k) + 2\sum_{k=1}^{K} q_5(k) - 2\sum_{k=1}^{K} q_6(k)$$

where

$$q_1(k) = \sum_{i:i \in \Omega_k, \delta_i = 1} \widehat{w}_i^2 \left\{ A_i - \frac{S_{1,k}(c_i)}{S_{0,k}(c_i)} \right\}^2 = \sum_{j=1}^{d(k)} \sum_{l:l \in D_j(k)} \widehat{w}_l^2 \left\{ A_l - \frac{S_{1,k}(j)}{S_{0,k}(j)} \right\}^2$$

$$= \sum_{j=1}^{d(k)} \left\{ 1 - \frac{S_{1,k}(j)}{S_{0,k}(j)} \right\}^2 \sum_{l:l \in D_j(k), A_l = 1} \widehat{w}_l^2 + \sum_{j=1}^{d(k)} \left\{ \frac{S_{1,k}(j)}{S_{0,k}(j)} \right\}^2 \sum_{l:l \in D_j(k), A_l = 0} \widehat{w}_l^2$$

$$q_2(k) = \sum_{i:i \in \Omega_k} \widehat{w}_i^2 A_i \exp\left(2A_i\hat{\theta}\right) \left\{ \sum_{l:\, \delta_l = 1} \frac{\widehat{w}_l I(i, l \in \Omega_k) I(T_l \le T_i)}{S_{0,k}(c_l)} \right\}^2$$

$$= \exp\left(2\hat{\theta}\right) \sum_{i:i \in \Omega_k, A_i = 1} \widehat{w}_i^2 \left\{ \sum_{l:\, \delta_l = 1} \frac{\widehat{w}_l I(i, l \in \Omega_k) I(T_l \le T_i)}{S_{0,k}(c_l)} \right\}^2$$

$$= \exp\left(2\hat{\theta}\right) \sum_{j=1}^{d(k)} \left\{ \sum_{l:l \in \mathcal{R}_j(k), A_l = 1} \widehat{w}_l^2 I(T_l < T_{k,\mathrm{j}+1}^D) \right\} \left\{ \sum_{r=1}^{j} \frac{\sum_{l:l \in D_r(k)} \widehat{w}_l}{S_{0,k}(r)} \right\}^2$$

$$q_3(k) = \sum_{i:i\in\Omega_k} \widehat{w}_i^2 \exp\left(2A_i\hat{\theta}\right)\left\{\sum_{l:\,\delta_l=1} \frac{\widehat{w}_l I(i,l\in\Omega_k)I(T_l\le T_i)S_{1,k}(c_l)}{S_{0,k}^2(c_l)}\right\}^2$$

$$= \exp\left(2\hat{\theta}\right)\sum_{i:i\in\Omega_k,A_i=1} \widehat{w}_i^2\left\{\sum_{l:\,\delta_l=1} \frac{\widehat{w}_l I(i,l\in\Omega_k)I(T_l\le T_i)S_{1,k}(c_l)}{S_{0,k}^2(c_l)}\right\}^2$$

$$+ \sum_{i:i\in\Omega_k,A_i=0} \widehat{w}_i^2\left\{\sum_{l:\,\delta_l=1} \frac{\widehat{w}_l I(i,l\in\Omega_k)I(T_l\le T_i)S_{1,k}(c_l)}{S_{0,k}^2(c_l)}\right\}^2$$

$$= \exp\left(2\hat{\theta}\right)\sum_{j=1}^{d(k)}\left\{\sum_{l:l\in\mathcal{R}_j(k),A_l=1} \widehat{w}_l^2\, I(T_l < T_{k,\mathrm{j}+1}^D)\right\}\left\{\sum_{r=1}^{j}\frac{\sum_{l:l\in D_r(k)}\widehat{w}_l S_{1,k}(r)}{S_{0,k}^2(r)}\right\}^2$$

$$+ \sum_{j=1}^{d(k)}\left\{\sum_{l:l\in\mathcal{R}_j(k),A_l=0} \widehat{w}_l^2\, I(T_l < T_{k,\mathrm{j}+1}^D)\right\}\left\{\sum_{r=1}^{j}\frac{\sum_{l:l\in D_r(k)}\widehat{w}_l S_{1,k}(r)}{S_{0,k}^2(r)}\right\}^2$$

$$q_4(k) = \sum_{i:i\in\Omega_k} \widehat{w}_i^2\delta_i A_i \exp\left(A_i\hat{\theta}\right)\left\{A_i - \frac{S_{1,k}(c_i)}{S_{0,k}(c_i)}\right\}\left\{\sum_{l:\,\delta_l=1} \frac{\widehat{w}_l I(i,l\in\Omega_k)I(T_l\le T_i)}{S_{0,k}(c_l)}\right\}$$

$$= \exp\left(\hat{\theta}\right)\sum_{i:i\in\Omega_k,A_i=1} \widehat{w}_i^2\delta_i\left\{1 - \frac{S_{1,k}(c_i)}{S_{0,k}(c_i)}\right\}\left\{\sum_{l:\,\delta_l=1} \frac{\widehat{w}_l I(i,l\in\Omega_k)I(T_l\le T_i)}{S_{0,k}(c_l)}\right\}$$

$$= \exp\left(\hat{\theta}\right)\sum_{j=1}^{d(k)}\left(\sum_{l:l\in D_j(k),A_l=1} \widehat{w}_l^2\right)\left\{1 - \frac{S_{1,k}(j)}{S_{0,k}(j)}\right\}\left\{\sum_{r=1}^{j}\frac{\sum_{l:l\in D_r(k)}\widehat{w}_l}{S_{0,k}(r)}\right\}$$

$$q_5(k) = \sum_{i:i\in\Omega_k} \widehat{w}_i^2 \delta_i \exp\left(A_i\hat{\theta}\right)\left\{A_i - \frac{S_{1,k}(c_i)}{S_{0,k}(c_i)}\right\}\left\{\sum_{l:\,\delta_l=1} \frac{\widehat{w}_l I(i,l\in\Omega_k)I(T_l\le T_i)S_{1,k}(c_l)}{S_{0,k}^2(c_l)}\right\}$$

$$= \exp\left(\hat{\theta}\right) \sum_{i:i\in\Omega_k,A_i=1} \widehat{w}_i^2 \delta_i \left\{1 - \frac{S_{1,k}(c_i)}{S_{0,k}(c_i)}\right\}\left\{\sum_{l:\,\delta_l=1} \frac{\widehat{w}_l I(i,l\in\Omega_k)I(T_l\le T_i)S_{1,k}(c_l)}{S_{0,k}^2(c_l)}\right\}$$

$$+ \sum_{i:i\in\Omega_k,A_i=0} \widehat{w}_i^2 \delta_i \left\{0 - \frac{S_{1,k}(c_i)}{S_{0,k}(c_i)}\right\}\left\{\sum_{l:\,\delta_l=1} \frac{\widehat{w}_l I(i,l\in\Omega_k)I(T_l\le T_i)S_{1,k}(c_l)}{S_{0,k}^2(c_l)}\right\}$$

$$= \exp\left(\hat{\theta}\right) \sum_{j=1}^{d(k)} \left(\sum_{l:l\in D_j(k),A_l=1} \widehat{w}_l^2\right)\left\{1 - \frac{S_{1,k}(j)}{S_{0,k}(j)}\right\}\left\{\sum_{r=1}^{j} \frac{\sum_{l:l\in D_r(k)} \widehat{w}_l\, S_{1,k}(r)}{S_{0,k}^2(r)}\right\}$$

$$+ \sum_{j=1}^{d(k)} \left(\sum_{l:l\in D_j(k),A_l=0} \widehat{w}_l^2\right)\left\{0 - \frac{S_{1,k}(j)}{S_{0,k}(j)}\right\}\left\{\sum_{r=1}^{j} \frac{\sum_{l:l\in D_r(k)} \widehat{w}_l\, S_{1,k}(r)}{S_{0,k}^2(r)}\right\}$$

$q_6(k)$

$$= \sum_{i:i\in\Omega_k} \widehat{w}_i^2 A_i \exp\left(2A_i\hat{\theta}\right)\left\{\sum_{l:\,\delta_l=1} \frac{\widehat{w}_l I(i,l\in\Omega_k)I(T_l\le T_i)}{S_{0,k}(c_l)}\right\}\left\{\sum_{l:\,\delta_l=1} \frac{\widehat{w}_l I(i,l\in\Omega_k)I(T_l\le T_i)S_{1,k}(c_l)}{S_{0,k}^2(c_l)}\right\}$$

$$= \exp\left(2\hat{\theta}\right) \sum_{i:i\in\Omega_k,A_i=1} \widehat{w}_i^2 \left\{\sum_{l:\,\delta_l=1} \frac{\widehat{w}_l I(i,l\in\Omega_k)I(T_l\le T_i)}{S_{0,k}(c_l)}\right\}\left\{\sum_{l:\,\delta_l=1} \frac{\widehat{w}_j I(i,l\in\Omega_k)I(T_l\le T_i)S_{1,k}(c_l)}{S_{0,k}^2(c_l)}\right\}$$

$$= \exp\left(2\hat{\theta}\right) \sum_{j=1}^{d(k)} \left\{\sum_{l:l\in\mathcal{R}_j(k),A_l=1} \widehat{w}_l^2\, I(T_l < T_{k,j+1}^D)\right\}\left\{\sum_{r=1}^{j} \frac{\sum_{l:l\in D_r(k)} \widehat{w}_l}{S_{0,k}(r)}\right\}\left\{\sum_{r=1}^{j} \frac{\sum_{l:l\in D_r(k)} \widehat{w}_l\, S_{1,k}(r)}{S_{0,k}^2(r)}\right\}$$

First, we observe that h can be preserved if $\sum_{l:l\in D_j(k),A_l=1}\widehat{w}_l$, $\sum_{l:l\in D_j(k)}\widehat{w}_l$, $\sum_{l:l\in\mathcal{R}_j(k),A_l=1}\widehat{w}_l$ and $\sum_{l:l\in\mathcal{R}_j(k),A_l=0}\widehat{w}_l$ are preserved. Second, we observe that q can be preserved if $S_{0,k}(j)$, $S_{1,k}(j)$,$\sum_{l:l\in D_j(k),A_l=1}\widehat{w}_l^2$, $\sum_{l:l\in D_j(k),A_l=0}\widehat{w}_l^2$, $\sum_{l:l\in\mathcal{R}_j(k),A_l=1}\widehat{w}_l^2 I(T_l < T_{k,\mathrm{j}+1}^D)$, $\sum_{l:l\in\mathcal{R}_j(k),A_l=0}\widehat{w}_l^2 I(T_l < T_{k,\mathrm{j}+1}^D)$, $\sum_{r=1}^{j}\frac{\sum_{l:l\in D_r(k)}\widehat{w}_l}{S_{0,k}(r)}$ and $\sum_{r=1}^{j}\frac{\sum_{l:l\in D_r(k)}\widehat{w}_l S_{1,k}(r)}{S_{0,k}^2(r)}$ are preserved.

Notice that $S_{0,k}(j)$, $S_{1,k}(j)$, $\sum_{r=1}^{j}\frac{\sum_{l:l\in D_r(k)}\widehat{w}_l}{S_{0,k}(r)}$ and $\sum_{r=1}^{j}\frac{\sum_{l:l\in D_r(k)}\widehat{w}_l S_{1,k}(r)}{S_{0,k}^2(r)}$ can be preserved if $\sum_{l:l\in D_j(k),A_l=1}\widehat{w}_l$, $\sum_{l:l\in D_j(k)}\widehat{w}_l$, $\sum_{l:l\in\mathcal{R}_j(k),A_l=1}\widehat{w}_l$ and $\sum_{l:l\in\mathcal{R}_j(k),A_l=0}\widehat{w}_l$ are preserved. We further note that

$$\sum_{l:l\in\mathcal{R}_j(k),A_l=1}\widehat{w}_l^2\, I\left(T_l < T_{k,\mathrm{j}+1}^D\right) = \sum_{l:l\in\mathcal{R}_j(k),A_l=1}\widehat{w}_l^2 - \sum_{l:l\in\mathcal{R}_j(k),A_l=1}\widehat{w}_l^2\, I\left(T_l \geq T_{k,\mathrm{j}+1}^D\right)$$

$$= \sum_{l:l\in\mathcal{R}_j(k),A_l=1}\widehat{w}_l^2 - \sum_{l:l\in\mathcal{R}_{j+1}(k),A_l=1}\widehat{w}_l^2$$

$$\sum_{l:l\in\mathcal{R}_j(k),A_l=0}\widehat{w}_l^2\, I\left(T_l < T_{k,\mathrm{j}+1}^D\right) = \sum_{l:l\in\mathcal{R}_j(k),A_l=0}\widehat{w}_l^2 - \sum_{l:l\in\mathcal{R}_j(k),A_l=0}\widehat{w}_l^2\, I\left(T_l \geq T_{k,\mathrm{j}+1}^D\right)$$

$$= \sum_{l:l\in\mathcal{R}_j(k),A_l=0}\widehat{w}_l^2 - \sum_{l:l\in\mathcal{R}_{j+1}(k),A_l=0}\widehat{w}_l^2$$

Therefore, $\sum_{l:l\in\mathcal{R}_j(k),A_l=1}\widehat{w}_l^2\, I\left(T_l < T_{k,\mathrm{j}+1}^D\right)$ and $\sum_{l:l\in\mathcal{R}_j(k),A_l=0}\widehat{w}_l^2\, I\left(T_l < T_{k,\mathrm{j}+1}^D\right)$ are preserved if $\sum_{l:l\in\mathcal{R}_j(k),A_l=1}\widehat{w}_l^2$ and $\sum_{l:l\in\mathcal{R}_j(k),A_l=0}\widehat{w}_l^2$ are preserved for each j.

In summary, the robust sandwich variance estimate will be the same as that from the corresponding pooled individual-level data analysis, if we preserve summary-level quantities

$\sum_{l:l\in D_j(k),A_l=1} \widehat{w}_l$ (total weights of treated events)

$\sum_{l:l\in D_j(k)} \widehat{w}_l$ (total weights of events)

$\sum_{l:l\in \mathcal{R}_j(k),A_l=1} \widehat{w}_l$ (total weights of treated patients)

$\sum_{l:l\in \mathcal{R}_j(k),A_l=0} \widehat{w}_l$ (total weights of untreated patients)

$\sum_{l:l\in D_j(k),A_l=1} \widehat{w}_l^2$ (total squared weights of treated events)

$\sum_{l:l\in D_j(k),A_l=0} \widehat{w}_l^2$ (total squared weights of untreated events)

$\sum_{l:l\in \mathcal{R}_j(k),A_l=1} \widehat{w}_l^2$ (total squared weights of treated patients)

and

$\sum_{l:l\in \mathcal{R}_j(k),A_l=0} \widehat{w}_l^2$ (total squared weights of untreated patients)

for the j^{th} ($j = 1, \dots, d(k)$) distinct observed event time in each site k. See Table 2 for an example of a risk-set dataset.

Appendix 2

The following figures show the steps required to estimate the hazard ratios and variances using the three proposed methods.

Figure A1. File transfers between the data-contributing sites and the analysis center for Method 1: Privacy-Protecting Estimation of Hazard Ratio and Robust Sandwich Variance

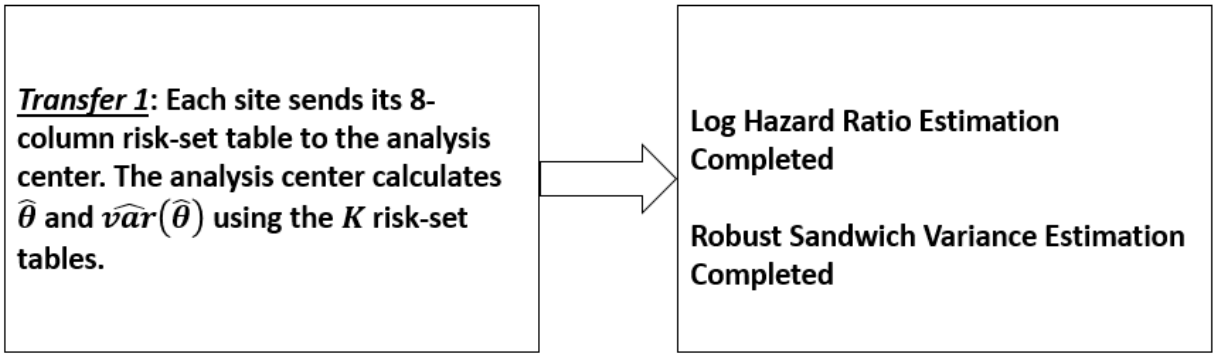


Figure A2. File transfers between the data-contributing sites and the analysis center for Method 2a: Privacy-Protecting Estimation of Hazard Ratio and Global Bootstrap Variance

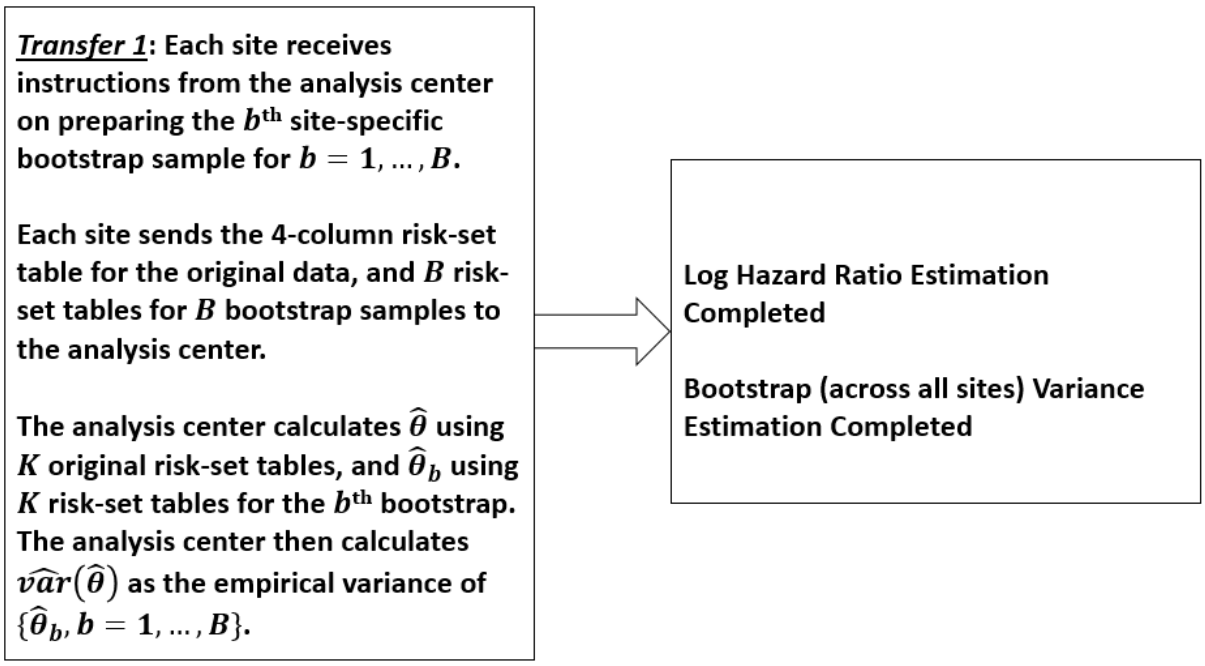
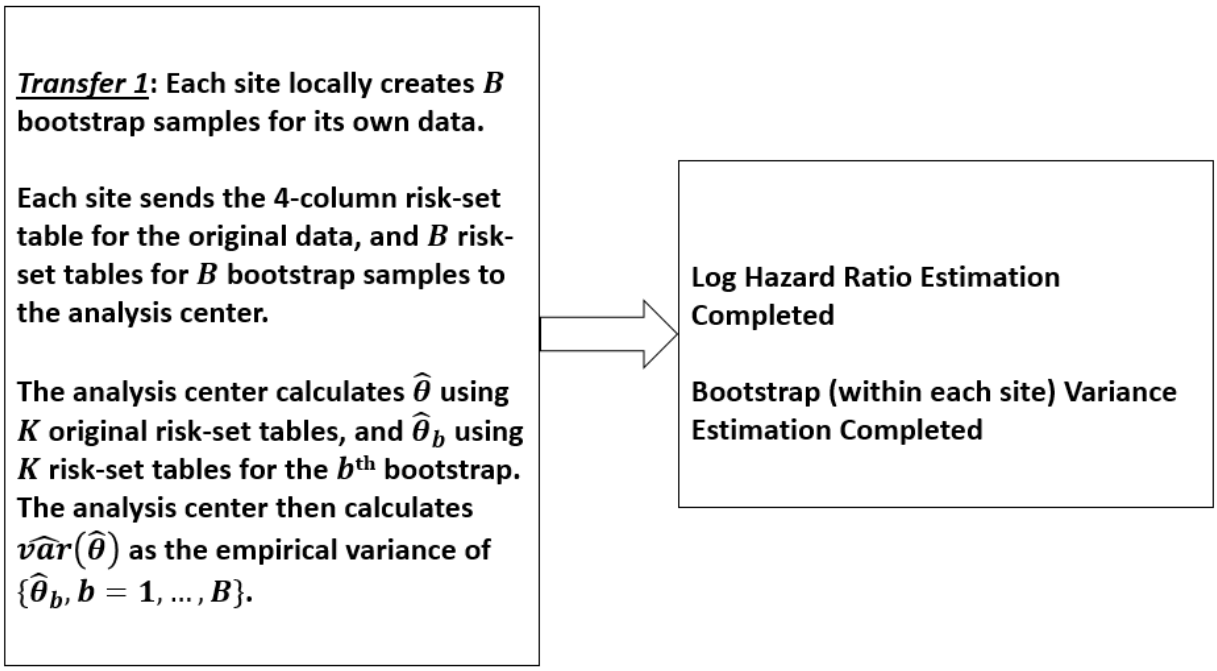


Figure A3. File transfers between the data-contributing sites and the analysis center for Method 2b: Privacy-Protecting Estimation of Hazard Ratio and Local Bootstrap Variance



Appendix 3

Section 0: R code for data generation

We supply the R code for data generation as described in Section 5.1.

Data with no tied events can be generated by code in the submitted file "*simulate_data.R*".

Data with tied events can be generated by code in the submitted file "*simulate_data_tied.R*".

Section 1: R code for Method 1: Privacy-protecting estimation of hazard ratio and robust sandwich variance

We supply the R code for two illustrative examples using Method 1 in the paper. The R analytic code for two illustrative examples is given in the submitted file "*Method1_simulated_data.R*".

Running the replication code in R will reproduce the log hazard ratio and standard error estimates reported in the paper. Comments are inserted in the R script for readability. In practice, users may have different number of sites or different number of covariates. The code here can be easily modified for specific questions.

Section 2: R code for Method 2a: Privacy-protecting estimation of hazard ratio and bootstrap variance with global bootstrap resampling

We supply the R code for two illustrative examples using Method 2a in the paper. The R analytic code for two illustrative examples is given in the submitted file "*Method2a_simulated_data.R*".

Running the replication code in R will reproduce the log hazard ratio and standard error estimates reported in the paper. Comments are inserted in the R script for readability. In practice, users may have different number of sites or different number of covariates. The code here can be easily modified for specific questions.

Section 3: R code for Method 2b: Privacy-protecting estimation of hazard ratio and bootstrap variance with local bootstrap resampling

We supply the R code for two illustrative examples using Method 2b in the paper. The R analytic code for two illustrative examples is given in the submitted file "*Method2b_simulated_data.R*".

Running the replication code in R will reproduce the log hazard ratio and standard error estimates reported in the paper. Comments are inserted in the R script for readability. In practice, users may have different number of sites or different number of covariates. The code here can be easily modified for specific questions.