eAppendix for

"ppmHR: A Privacy-Protecting Tool to Fit Inverse Probability Weighted Cox Models in Multi-Site Studies"

Di Shu and Sengwee Toh

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

Part A: algorithm for fitting inverse probability weighted Cox models from risk set tables

Step 0: Re-organize the shared risk set tables such that the j^{th} row represents the j^{th} ranked event times. This can be done by sorting the tables such that columns sumSquareE and sumSquareUnE are both non-increasing.

Point estimation:

Let k indicate the k^{th} site. The estimated hazard ratio (HR) can be obtained by solving the estimating equation

$$\sum_k y_k = 0$$

where $y_k = \sum\left\{sumEC(k) - sumC(k)\frac{sumE(k)*HR}{sumE(k)*HR+sumUnE(k)}\right\}$, and the outside summation means adding up all elements in vector $\left\{sumEC(k) - sumC(k)\frac{sumE(k)*HR}{sumE(k)*HR+sumUnE(k)}\right\}$.

Step 1: Since $sumEC(k)$, $sumC(k)$, $sumE(k)$, and $sumUnE(k)$ are available as the 1st to 4th columns of the shared risk set tables, we can write code to use existing root-solving programs (e.g., nleqslv package in R) to solve the above equation for HR.

Robust sandwich variance estimation:

The robust sandwich variance estimator of the estimated log HR, log(HR), is given by q/h^2. Below we calculate h and q separately.

Step 2: Calculate $S_0(k)$ and $S_1(k)$, where

$$S_0(k) = sumE(k) * HR + sumUnE(k)$$

$$S_1(k) = sumE(k) * HR$$

and HR is the point estimate obtained from Step 1.

Step 3: Calculate $h = \sum_k h_k$ where

$$h_k = \sum sumC(k)\left\{\frac{S_1(k)}{S_0(k)} - \left(\frac{S_1(k)}{S_0(k)}\right)^2\right\}$$

Step 4: Calculate $sumSquEdiff(k) = \text{sumSquareE(k)} - \text{sumSquareE\_lag(k)}$, where sumSquareE_lag(k) is sumSquareE(k) starting from the second element and having 0 as the last element. For example, if the column vector sumSquareE(k) is (1,2,3,4), then the column vector sumSquareE_lag(k) is (2,3,4,0).

Step 5: Calculate $sumSquUnEdiff(k) = \text{sumSquareUnE(k)} - \text{sumSquareUnE\_lag(k)}$, where $\text{sumSquareUnE\_lag(k)}$ is sumSquareUnE(k) starting from the second element and having 0 as the last element.

Step 6: Calculate $cumsum1(k) = cumsum(sumC/S_0(k))$, where cumsum is the cumulative summation operator. For example, for column vector a=(1,2,3), cumsum(a)=(1,3,6).

Step 7: Calculate $cumsum2(k) = cumsum(sumC * \frac{S_1(k)}{S_0^2(k)})$, where cumsum is cumulative summation operator.

Note that sumSquareE(k) and sumSquareUnE(k) are available as the 7th and 8th columns of the shared risk set tables, so Steps 4 to 7 can be done.

Step 8: Calculate $q_1(k)$ to $q_6(k)$ where

$$q_1(k) = \sum \left\{ sumSquareEC(k) \left(1 - \frac{S_1(k)}{S_0(k)}\right)^2 + sumSquareUnEC(k) \left(\frac{S_1(k)}{S_0(k)}\right)^2 \right\}$$

$$q_2(k) = \exp(2logHR) \sum \{sumSquEdiff(k) * cumsum1(k)^2\}$$

$$q_3(k) = \exp(2logHR) \sum \{sumSquEdiff(k) * cumsum2(k)^2\}$$
$$+ \sum \{sumSquUnEdiff(k) * cumsum2(k)^2\}$$

$$q_4(k) = \mathrm{HR} \sum \left\{ sumSquareEC(k) * \left(1 - \frac{S_1(k)}{S_0(k)}\right) * cumsum1(k) \right\}$$

$$q_5(k) = \mathrm{HR} \sum \left\{ sumSquareEC(k) * \left(1 - \frac{S_1(k)}{S_0(k)}\right) * cumsum2(k) \right\}$$
$$+ \sum \left\{ sumSquareUnEC(k) * \left(0 - \frac{S_1(k)}{S_0(k)}\right) * cumsum2(k) \right\}$$

$$q_6(k) = \exp(2logHR) \sum \{sumSquEdiff(k) * cumsum1(k) * cumsum2(k)\}$$

Note that $sumSquareEC(k)$ and $sumSquareUnEC(k)$ are available as the 5th and 6th columns of the shared risk set tables, and $sumSquEdiff(k)$, $sumSquUnEdiff(k)$, $cumsum1(k)$, and $cumsum2(k)$ are obtained from Steps 4-7, so Step 8 can be done.

Step 9: Calculate

$$q = \sum_k q_1(k) + \sum_k q_2(k) + \sum_k q_3(k) - 2\sum_k q_4(k) + 2\sum_k q_5(k) - 2\sum_k q_6(k)$$

Step 10: Calculate the robust sandwich variance estimate q/h^2.

Part B: Replication R code

```
library(ppmHR)
#load example datasets in the package
#site1-3 contain individual-level data of data-contributing sites 1-3
data(site1)
data(site2)
data(site3)
#data-contributing sites 1-3 create summary-level risk set tables
#with logistic propensity score model A~X1+X2+X3+X4+X5
#no weight truncation
rsTb1=createRisksetTable(data=site1,indA="A",indX=c("X1","X2","X3","X4","X5"),
                indStatus="status",indTime="time")
rsTb2=createRisksetTable(data=site2,indA="A",indX=c("X1","X2","X3","X4","X5"),
                indStatus="status",indTime="time")
rsTb3=createRisksetTable(data=site3,indA="A",indX=c("X1","X2","X3","X4","X5"),
                indStatus="status",indTime="time")
#analysis center estimates overall hazard ratios in a stratified IPW Cox model
#using summary-level risk set tables rsTb1-3 shared by data-contributing sites
estimateStratHR(list(rsTb1,rsTb2,rsTb3))

#analysis center estimates site-specific hazard ratios in IPW Cox models
estimateStratHR(list(rsTb1))
estimateStratHR(list(rsTb2))
estimateStratHR(list(rsTb3))

#pooled individual-level data analysis using survival package
library(survival)
#calculate weights
psmd=glm(A~X1+X2+X3+X4+X5,data=site1,family=binomial(link='logit'))
psfit=predict(psmd,type = "response",data=site1)
site1$wt=site1$A/psfit+(1-site1$A)/(1-psfit)
site1$swt=mean(site1$A)*site1$A/psfit+mean(1-site1$A)*(1-site1$A)/(1-psfit)

psmd=glm(A~X1+X2+X3+X4+X5,data=site2,family=binomial(link='logit'))
psfit=predict(psmd,type = "response",data=site2)
site2$wt=site2$A/psfit+(1-site2$A)/(1-psfit)
site2$swt=mean(site2$A)*site2$A/psfit+mean(1-site2$A)*(1-site2$A)/(1-psfit)

psmd=glm(A~X1+X2+X3+X4+X5,data=site3,family=binomial(link='logit'))
psfit=predict(psmd,type = "response",data=site3)
site3$wt=site3$A/psfit+(1-site3$A)/(1-psfit)
site3$swt=mean(site3$A)*site3$A/psfit+mean(1-site3$A)*(1-site3$A)/(1-psfit)

#pooled individual-level data analysis for overall hazard ratios
```

```
pooled=rbind(site1,site2,site3)
fit <- coxph(Surv(time, status) ~ A+cluster(id)+strata(indSite),
weights=wt,data=pooled, ties="breslow")
c(fit$coefficients,fit$var^0.5)

fits <- coxph(Surv(time, status) ~ A+cluster(id)+strata(indSite),
weights=swt,data=pooled, ties="breslow")
c(fits$coefficients,fits$var^0.5)

#pooled individual-level data analysis for site-specific hazard ratios
#site 1
fit <- coxph(Surv(time, status) ~ A+cluster(id), weights=wt,data=site1,
ties="breslow")
c(fit$coefficients,fit$var^0.5)
fits <- coxph(Surv(time, status) ~ A+cluster(id), weights=swt,data=site1,
ties="breslow")
c(fits$coefficients,fits$var^0.5)

#site2
fit <- coxph(Surv(time, status) ~ A+cluster(id), weights=wt,data=site2,
ties="breslow")
c(fit$coefficients,fit$var^0.5)
fits <- coxph(Surv(time, status) ~ A+cluster(id), weights=swt,data=site2,
ties="breslow")
c(fits$coefficients,fits$var^0.5)

#site3
fit <- coxph(Surv(time, status) ~ A+cluster(id), weights=wt,data=site3,
ties="breslow")
c(fit$coefficients,fit$var^0.5)
fits <- coxph(Surv(time, status) ~ A+cluster(id), weights=swt,data=site3,
ties="breslow")
c(fits$coefficients,fits$var^0.5)
```