SBspecs080116.doc

Methods for Adjusting for Cohort Selection Bias: Simulations to Evaluate Proposed Methods

1. Overview.

Our goal is to use simulations to evaluate a proposed weighting method for addressing covariate selection bias in cohorts of subjects followed for occurrence of a given outcome. The covariates of subjects who participate in cohort studies are seldom representative of those of the general population, and such bias can give misleading estimates of the performances of personal risk models that use subjects' covariates to assign them personalized probabilities of developing future adverse outcomes. The weighting method works by supplementing the covariate data of the cohort subjects with cross-sectional covariate data from a population-based sample of subjects whose covariate distribution better represents that of the population for whom the personal predictive model (PPM) of interest in intended. Specifically, we use the covariates in the PPM to assign risks to subjects in the cohort and in a sample of the target population, and compare the two risk distributions. If they differ appreciably, we weight the risks of cohort subjects so that their distribution more closely matches that of the population risks, and then evaluate performance metrics using the weighted distribution of cohort risks. This approach assumes that, conditional on the covariates used by the risk model, the outcome probabilities of cohort subjects represent those of the target population.

To assess differences in the distributions of PPM-assigned risks among subjects in the cohort and population-based samples, we model both sets of assigned risks as beta distributions of the form

$$g(r)=\left[B(\alpha,\beta)\right]^{-1} r^{\alpha-1}(1-r)^{\beta-1},\ 0<r<1,$$

where $B(\alpha,\beta)=\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\,du$ is the beta function. The full likelihood of the combined data is then

$$L(\alpha_C,\beta_C,\alpha_P,\beta_P)=B(\alpha_C,\beta_C)^{-N_C}\,B(\alpha_P,\beta_P)^{-N_P}\prod_{i=1}^{N_C} r_{Ci}^{w_{Ci}(\alpha_C-1)}(1-r_{Ci})^{w_{Ci}(\beta_C-1)}\prod_{i'=1}^{N_P} r_{Pi'}{}^{w_{Pi'}(\alpha_P-1)}(1-r_{Pi'})^{w_{Pi'}(\beta_P-1)}\quad,$$

with loglikelihood

$$\begin{aligned}\ell(\alpha_C,\beta_C,\alpha_P,\beta_P)&=(\alpha_C-1)\sum_{i=1}^{N_C} w_{Ci}\ln r_{Ci}+(\beta_C-1)\sum_{i=1}^{N_C} w_{Ci}\ln(1-r_{Ci})-N_C\ln\left[B(\alpha_C,\beta_C)\right]\\&\quad+(\alpha_P-1)\sum_{i'=1}^{N_P} w_{Pi'}\ln r_{Pi'}+(\beta_P-1)\sum_{i'=1}^{N_P} w_{Pi'}\ln(1-r_{Pi'})-N_P\ln\left[B(\alpha_P,\beta_P)\right]\end{aligned}$$

Here r_{Ci} and $r_{Pi'}$ denote the risks assigned to sampled subjects from C and P, respectively, and $w_{Ci} = 1,\ i = 1,\ldots,N_C$. The loglikelihood of the data under H_0 is

$$\begin{aligned}\ell(\alpha,\beta)&=(\alpha-1)\left\{\sum_{i=1}^{N_C} w_{Ci}\ln r_{Ci}+\sum_{i'=1}^{N_P} w_{Pi'}\ln r_{Pi'}\right\}\\&\quad+(\beta-1)\left\{\sum_{i=1}^{N_C} w_{Ci}\ln(1-r_{Ci})+\sum_{i'=1}^{N_P} w_{Pi'}\ln(1-r_{Pi'})\right\}-(N_C+N_P)\ln\left[B(\alpha,\beta)\right]\end{aligned}.$$

We obtain maximum likelihood estimates (MLEs) of the parameters in the full and null loglikehoods, using as initial estimates the parameter values obtained with the method of moments. Specifically, we took $\hat{\alpha}^0=\bar{r}a$ & $\hat{\beta}^0=(1-\bar{r})a$, where $a=\left[\bar{r}(1-\bar{r})/s^2\right]-1$ and $\bar{r}$ & s^2 are the sample mean and variance, given by

$$\bar{r}_C=N_C^{-1}\sum_{i=1}^{N_C} w_{Ci}r_{Ci}\ \&\ s_c^2=(N_C-1)^{-1}\sum_{i=1}^{N_C} w_{Ci}(r_{Ci}-\bar{r}_C)^2$$

$$\bar{r}_P=N_P^{-1}\sum_{i'=1}^{N_P} w_{Pi'}r_{Pi'}\ \&\ s_P^2=(N_P-1)^{-1}\sum_{i'=1}^{N_P} w_{Pi'}(r_{Pi'}-\bar{r}_P)^2$$

$$\text{and } \bar{r}=(N_C+N_P)^{-1}(N_C\bar{r}_C+N_P\bar{r}_P)\ \&\ s^2=(N_C+N_P-1)^{-1}\left[\sum_{i=1}^{N_c} w_{Ci}(r_{Ci}-\bar{r})^2+\sum_{i'=1}^{N_P} w_{Pi'}(r_{Pi'}-\bar{r})^2\right]$$

If the covariate distributions in cohort and population differ, we evaluate PPM performance using cohort subjects who have been weighted to make their covariate distribution more similar to that of the population.

To obtain these weights, we classify the subjects in each of the two samples into *J* joint covariate categories, with $\hat{\varphi}_{Cj}$ and $\hat{\varphi}_{Pj}$ denoting, respectively, the estimated proportions of cohort & target population in category *j*, *j=1, …,,J*. Then we assign the ith cohort subject the weight

$$w_i = \sum_{j=1}^{J} \frac{\hat{\varphi}_{Pj}}{\hat{\varphi}_{Cj}} 1\left(i \in cat\ j\right),\ i = 1,...,N_C. \tag{3}$$

When $\hat{\varphi}_{Cj} = N_C^{-1} \sum_{i=1}^{N_C} 1\left(i \in \text{ cat } j\right)$, the sum of the weights over all cohort subjects is *N_C.* If the null hypothesis of equal covariate distributions is not rejected by the weighted test, we evaluate PPM performance by weighting the cohort subjects' survival data.

2. Evaluating PPM Performance.

Model Calibration. We evaluate the calibration of a PPM to a cohort by comparing the subjects' empirical outcome probabilities to their mean assigned risks. Specifically, we partition the cohort subjects into *L* subgroups based on their PPM-assigned risks, and evaluate the GOF statistics

$$X_1^2 = \frac{\left(\hat{\pi} - \overline{r}\right)^2}{\widehat{\text{var}}\left(\hat{\pi}\right)} \text{ and } X_L^2 = \sum_{\ell=1}^{L} \hat{\gamma}_\ell \frac{\left(\hat{\pi}_\ell - \overline{r}_\ell\right)^2}{\widehat{\text{var}}\left(\hat{\pi}_\ell\right)}. \tag{9}$$

Here $\overline{r}_\ell$ is the (wtd/unwtd) mean assigned risk and $\hat{\gamma}_\ell$ is the (wtd/unwtd) proportion of subjects in risk-group $\ell,\ \ell = 1,...,L$. We regard the $\hat{\gamma}_\ell$ as fixed in determining the bootstrap variance estimates $\widehat{\text{var}}\left(\hat{\pi}_\ell\right)$, conditional on the observed values $\hat{\gamma}_1,...,\hat{\gamma}_L$. (See Appendix A). Given $\hat{\gamma}_1,...,\hat{\gamma}_L,\ X_L^2$ is a quadratic form in *L* Gaussian variables, whose asymptotic null distribution is that of a mixture of central chi-squared distributions: $X_L^2 \sim \sum_{\ell=1}^{L} \hat{\gamma}_\ell \chi_{\ell 1}^2$. Several approximations to this distribution have been proposed[4]. The exact method of Davies[5] has been found to perform well and can be implemented in an R-package (see attached description.)

Model Discrimination. We evaluate model discrimination by estimating the concordance statistic (also called the AUC), which is the probability that the assigned risk for a randomly chosen outcome-positive subject (one who develops the outcome in the period (0,t*)) exceeds that of a randomly sampled outcome-negative subject (one who is alive and outcome-free at time t* = 1). Estimating C with censored survival data is problematic, because subjects last observed outcome-free before t* are outcome-unknown. Moreover the estimate obtained by excluding these subjects is inefficient and biased upward, even when censoring times are independent of outcome times and of assigned risk *(Melcon et al 2016)*. An alternative *Bernoulli* estimate, denoted as $\widehat{AUC}_{Brn}$, (Melcon et al 2016), described in Appendix C, exhibits systematically smaller bias and variance than the deletion-based estimate obtained by excluding all censored subjects. $\widehat{AUC}_{Brn}$ is obtained by generating for each outcome-unknown subject a random pseudo-outcome indicator whose Bernoulli probability depends on the subject's available survival information and model-assigned risk. The AUC is then estimated using all subjects' data, with pseudo-outcomes replacing known outcomes as needed. Specifically,

$$\widehat{AUC} = \sum_{r=0}^{1} \hat{h}_1\left(r\right)\hat{H}_0\left(r\right), \text{ where } \hat{h}_y\left(r\right) = \left[\sum_{i:y_i=y} w_i 1\left(r_i = r\right)\right] / \left[\sum_{i:y_i=y} w_i\right], y = 0,1 \ , \tag{10}$$

denotes the empirical pdf of risks among outcome-positive (y=1) and outcome-negative (y=0) subjects, and $\hat{H}_y\left(r\right)$ is the corresponding cdf. We obtain variance estimates and corresponding 95% CI's by bootstrapping the data, as described in Appendix B.

3. **Simulation**.

Overview. We now describe a simulation to generate cross-sectional covariate data for population-based and cohort samples, plus survival data for the cohort subjects, and use the data (with & without cohort weighting) to evaluate the performance of a hypothetical PPM. PPM performance is evaluated using two criteria: 1) calibration (agreement between predicted and observed outcome probabilities) and discrimination (extent of separation between the distributions of model-assigned risk among outcome-positive and outcome-negative subjects). For calibration, we will divide each sample into *L = 4* risk groups determined by fixed cutpoints of assigned risk, and compare observed & predicted outcome probabilities. For discrimination, we will examine the area under the receiver operating characteristic curve (AUC). The simulation involves E = 1000 experiments; in each experiment we generate and analyze covariate and survival data for three cohorts and cross-sectional covariate data for an additional population sample for use in weighting the cohort subjects.

3.1. Data generation for one experiment.

Generating subjects' covariates. We assume two covariates *(z_1,z_2)* having a Gaussian distribution in the target population P with mean $\left(\mu_1,\mu_2\right)=\left(-2.50,0.50\right)$ and covariance $V=\operatorname{diag}\left(\sigma_1^2,\sigma_2^2\right)=\operatorname{diag}\left(0.640,0.562\right)$. We generate covariate data for three cohorts and for an additional cross-sectional population sample for use in weighting cohort subjects. Covariates for subjects in cohorts *C_1* and *C2* are obtained by oversampling small values of z1 and z2, while those in C3 are obtained by simple random sampling. The sampling design is shown in Table A.1.

Assigning PPM risks. We assume a negligible competing risk of dying without the outcome during the risk period, and base the simulation on risk models of the form

$$r_i = 1-e^{-\lambda_{i1}},\ i=1,\ldots,N,\ \text{where } \lambda_{i1}=e^{z_{i1}+z_{i2}} \qquad (4)$$

is the (time-independent) outcome hazard rate. Equation (4) specifies the probability of developing the outcome within the risk period (0,t*), with t* = 1. We consider PPMs:

$$\text{Model A: } \lambda_{i1}=e^{z_{i1}+z_{i2}};\ \text{Model B: } \lambda_{i1}=e^{z_{i1}}. \qquad (5)$$

Note from (5) that Model A correctly specifies the outcome hazards, while Model B ignores the covariate *z_{i2}*.

Generating censored survival data. The *i*th subject's times $t_{i\tau}$ to censoring ($\tau=0$) & outcome (τ = 1), given his/her covariates, are generated according to independent exponential density functions $t_{i\tau}\sim f_{i\tau}\left(t\right)=\lambda_{i\tau}e^{-\lambda_{i\tau}t}$ with hazard parameters $\lambda_{i\tau},\tau=0,1$. We take a common value $\lambda_{i0}\equiv 0.056$ for the censoring hazard, for all subjects. We take the outcome hazard as $\lambda_{i1}=e^{z_{i1}+z_{i2}}$. We record each cohort subject's observed data as $\left(t_i,\varepsilon_i\right)$, where *$t_i$ = min(t_{i0},t_{i1}, t*=1)* and ε_i takes value 1 if *t_i* = t_{i1} and zero otherwise.

3.2 Data analysis for one experiment.

Once we have generated covariates and model risks for samples *C_1-C_3*, and *P,* and generated survival data for samples *C_1-C_3*, we estimate the model performance measures described in Section 2. To do so, we obtain unweighted and weighted estimates using data from one of the 2 PPMs and 3 cohorts (i.e., 2X3 = 6 pairs of wtd/unwtd estimates). The unweighted estimates are obtained by setting *w_{Ci} = 1* for all subjects, and the weighted ones by classifying the subjects in both population and cohort samples into J = 4 covariate categories shown in Table A.1. and using formula (3). Since the bootstrap variance estimates are larger for weighted than unweighted analyses (see Appendix A), comparison of the weighted & unweighted results for the unbiased cohort *C_3* will tell us how much increased uncertainty is caused by needless weighting.

Estimating outcome probabilities. To evaluate calibration, we compare PPM-assigned risks with outcome probabilities estimated using the sampled cohort survival data. For example, we compare these quantities for subjects in each of *L = 4* risk-groups:

$$S_\ell=\left\{i: a_{\ell-1}\le r_i<a_\ell\right\},\ell=1,\ldots 4,\ \text{with } (a_0,a_1,\cdots,a_4)=\left(0.0,0.10,0.15,0.20,1.0\right)$$

We want to estimate the overall outcome probability π among all subjects in a sample, and the risk-group-specific outcome probabilities $\pi_1,...,\pi_4$, where $\pi_\ell = 1 - e^{-\lambda_\ell}$ is the mean outcome probability at time *t** = 1 among subjects in $S_\ell, \ell = 1,...,L.$ We estimate the π_l by obtaining NPMLEs of the outcome hazards λ_ℓ.[1-3] Specifically, let M_ℓ denote the number of distinct event times in subgroup *l*, with $M = \sum_{\ell=1}^{4} M_\ell$. The risk-group- and event-specific hazard functions are replaced by the *M*-dimensional vector $\lambda = (\lambda_1,...,\lambda_5)$, where

$$\lambda_\ell = \left(\lambda_{\ell 1},...,\lambda_{\ell M_l}\right) \tag{6}$$

is the vector of discrete hazards taking values at the M_ℓ distinct event times $0 < t_{\ell 1} < \cdots < t_{\ell M_\ell} < t_*$ among subjects in group ℓ, $\ell = 1,...,L.$ It can be shown that the NPMLE is $\hat{\lambda} = \left(\hat{\lambda}_1,...,\hat{\lambda}_L\right)$, where $\hat{\lambda}_\ell$ is given by (6) with its elements $\hat{\lambda}_{\ell m}$ computed as follows. First define for each subject the left-continuous "at-risk" counting process $Y_i(t) = 1(t \le t_i)$, $i = 1,...,N$, where $1(E)$ is the indicator function taking value 1 if event *E* is true and zero otherwise. Also define the right-continuous "event" counting process $N_i(t) = 1(t \ge t_i, \varepsilon_i = 1)$. Then $\hat{\lambda}_{\ell m} = d_{\ell m} / n_{\ell m}$, where

$$n_{\ell m} = \sum\nolimits_{i \in Q_\ell} w_i Y_i(t_{\ell m}) \;\&\; d_{\ell m} = \sum\nolimits_{i \in Q_\ell} w_i Y_i(t_{\ell m}) N_i(t_{\ell m}),\; m = 1,...,M_\ell,\; \ell = 1,...,L. \tag{7}$$

Thus $n_{\ell m}$ and $d_{\ell m}$ are the weighted counts of subjects in risk-group S_ℓ who at time $t_{\ell m}$ are at risk and fail, respectively. We now estimate π_ℓ as

$$\hat{\pi}_\ell = \varphi_\ell\left(\hat{\lambda}_\ell\right) = \sum\nolimits_{m=1}^{M_\ell} \hat{\lambda}_{\ell m} \prod\nolimits_{m'=1}^{m-1}\left(1 - \hat{\lambda}_{\ell m}\right),\; \ell = 1,...,L, \tag{8}$$

where empty products equal one. In the absence of censoring, $\hat{\pi}_l$ reduces to the simple binomial proportion of subjects in risk-group ℓ who develop the outcome before dying during the risk period. The overall estimate $\hat{\pi}$ is obtained similarly.

Simulation Question: To better match the simulation design to that of the CTS ovarian cancer example, should we make either of the following changes?

a) oversample subjects with large covariate values z_1,z_2 (rather than small values). This would bias the cohort toward higher PPM risks, in agreement with the higher risk of the CTS cohort;

b) increase the cohort size to N_C = 10,000 and decrease the population sample size to N_P = 5000. This would better match the data application (*N_C = 46K, N_P = 2K*).

Tasks for E =1 Experiment.

A. Comparing weighted and unweighted distributions of assigned risk.

A.1. Create tables giving the empirical distributions of subjects in cohorts *C_1-C_3* across the two covariate categories (*z_{i2}<0,z_{i2}>=0*), and the corresponding values of the sampling weights.

A.2. Plot unweighted and weighted box plots of the empirical pdfs of assigned risks for the two risk models (A,B) and three cohort samples (*C_1-C_3*).

A.3. Use weighted and unweighted versions of the LR test statistic to compare each of the three cohort risk distributions to the population-based distribution (6 tests).

B. Calibration.

For each cohort and risk model:

B.1. obtain weighted and unweighted estimates and 95% CIs for the overall and risk-group-specific observed outcome probabilities (see Appendix B for details).

B.2 Create plots of points (x,y) where x denotes risk-group-specific mean assigned risks and y denotes risk-group-specific observed probability (with 95% confidence bars); include the diagonal for reference.

B.3. Use R software to calculate the value and significance level of the GOF statistic X_L^2 of equation (9).

C. Discrimination.

C.1 For each of the 6 cohort/PPM combinations, use the Bernoulli-based estimator described in Appendix C to obtain weighted and unweighted estimates and 95% CIs for the AUC. (See Appendix B.)

REFERENCES

1. Andersen PK, Borgan O, Gill RD, Keiding N. Statistical models based on counting processes. New York: Springer Verlag, 1992.
2. Dinse GE, Larson MG. A note on semi-Markov models for partially censored data. Biometrika 1986; 73(2): 379-386.
3. Kalbfleisch JD, Prentice RL. The Statistical Analysis of Failure Time Data, 2nd edition. New York: John Wiley & Sons, 2002.
4. Duchesne P and Lafaye de Micheaux. (2010) Comput. Stat. Data Anal. 54: 858-862.
5. Davies RB. (1980) the distribution of a linear combination of χ^2 random variables. JRSS Series C Applied Stats. 29: 323-333.
6. Melcon E et al. Estimating the concordance of prediction models with censored survival data (*submitted*).

APPENDIX

A. Bootstrap variance estimates. We bootstrap differently for weighted and unweighted statistics. Here are the steps for a given bootstrap replication b, b = 1,…,B.

Unweighted statistics. a) Randomly sample with replacement N_C cohort subjects; b) use their assigned risks and survival data to compute estimates for the statistics of interest. (For the GOF statistic of equation (9), fix the coefficients at their observed values $\hat{\gamma}_1,...,\hat{\gamma}_L$.)

Weighted statistics. Proceed in two steps: in Step 1, randomly sample with replacement N_P population-based subjects AND N_C subjects in each of the three cohorts, & classify each sample to get weights for the cohort subjects. In step 2, use the weights to compute estimates for the statistics of interest. (For the GOF statistic of equation (9), fix the coefficients at their observed values $\hat{\gamma}_1,...,\hat{\gamma}_L$.)

B. 95% CIs for observed performance metrics. The estimated outcome probabilities and AUC are all confined to the unit interval. To increase the accuracy of CIs and avoid bounds outside the unit interval, we obtain CIs for the logit transform of the estimate, get its confidence bounds, and then de-transform them with the inverse (expit) transform. Specifically, let y = logit x = log[x/(1-x)] and take SD(y) = SD(x)/[x(1-x)]. Then a 95% CI for x is given by

$$\left(\frac{e^{y_L}}{1+e^{y_L}},\frac{e^{y_U}}{1+e^{y_U}}\right), \text{ where } y_L = y-1.96SD(y) \text{ and } y_U = y+1.96SD(y), \text{ with } SD(y)=\frac{SD(x)}{x(1-x)}.$$

C. Bernoulli-based AUC estimates. The estimates $\widehat{AUC}_{Brn}$ are obtained by generating for each outcome-unknown subject *i* a pseudo-outcome $\hat{y}_i$ whose values $\hat{y}_i = 1$ and $\hat{y}_i = 0$ indicate outcome positivity and negativity, respectively. Each $\hat{y}_i$ is generated using the PPM-assigned conditional probability P_i that subject *i* develops the outcome by time t*, given her outcome-free survival until her last observation time t_i, $0 < t_i < t^*$. For the simulation, we take

$$P_i = 1-\Pr\left(\tilde{T}_i > t^* \mid \tilde{T}_i > t_i\right) = 1-\left[e^{\Lambda_i(t_i)-\Lambda_i(t^*)}\right], \qquad (*)$$

where $\Lambda_i(t)$ denotes her model-assigned cumulative hazard at time t.

For the data application, CTS subjects' PPM-assigned cumulative hazards are not available from the PPM software, so we approximate them by linearly interpolating between the available values $\Lambda_i(0)=0$ at time t = 0 and $\Lambda_i(t^*)=-\ln(1-r_i)$ at time t = t*, where r_i denotes the subject's PPM-assigned probability of outcome development by time *t**. Thus we took $P_i = 1-(1-r_i)^{1-t_i/t^*}$.

We used the bootstrap to obtain variance estimates and corresponding confidence intervals (see Appendix B) for these AUC estimates.