Web-based Supplementary Materials for Combining Information from Cancer Registry and Medical Records Data to Improve Analyses of Adjuvant Cancer Therapies

by Yulei He and Alan M. Zaslavsky

1 Gibbs Sampling Algorithm

We only include random intercepts in both the outcome and reporting domains of the latent variable models (1) and (2), with $L = 2$. Suppose S contains m hospitals, and in hospital i there are n_i patients. The observed data consist of $\{Y_{O1ij}^{obs}\}$ and $\{Y_{O2ij}^{obs}\}$ in S_1, $\{Y_{R1ij}\}$ and $\{Y_{R2ij}\}$ in S, and covariates $\{\mathbf{X}_{Oij}\}$ and $\{\mathbf{X}_{Rij}\}$ in S. The Gibbs sampler must draw

a. Latent variables $\{Z_{O1ij}\}$, $\{Z_{O2ij}\}$, $\{Z_{R1ij}\}$, and $\{Z_{R2ij}\}$;

b. Parameters: $\boldsymbol{\beta}_O$, $\boldsymbol{\beta}_R$, $\boldsymbol{\alpha} = (\alpha_{21}, \alpha_{12})^T$, $\boldsymbol{\rho}_O = \begin{pmatrix} 1 & \rho_O \\ \rho_O & 1 \end{pmatrix}$, $\boldsymbol{\rho}_R = \begin{pmatrix} 1 & \rho_R \\ \rho_R & 1 \end{pmatrix}$, and $\boldsymbol{\Sigma}$;

c. Random effects $\boldsymbol{\gamma}_{Oi}$ and $\boldsymbol{\gamma}_{Ri}$, $i = 1, \ldots, m$;

d. Missing values $\{Y_{O1ij}^{mis}\}$ and $\{Y_{O2ij}^{mis}\}$ in S_2.

When Y_{Olij} is observed or imputed to be 0, then $Y_{Rlij} = 0$ regardless of the value of Z_{Rlij}; hence we need not draw Z_{Rlij}. This reduces the amount of latent data and also reduces serial dependence in draws of reporting model parameters $\boldsymbol{\beta}_R$ and $\{\boldsymbol{\gamma}_{Ri}\}$. Define subsets of subjects $SC_{ab} = \{(i,j) : Y_{O1ij} = a, Y_{O2ij} = b\}, a, b = 0, 1$. As specified in Section 2.4, the sampling steps for the latent variables and parameters of the reporting model are only conducted in S_1.

The detailed steps of the algorithm are as follows:

Step 1. Draw latent variables $\{Z_{O1ij}, Z_{O2ij}\}$ from truncated bivariate normal distributions with mean $\mathbf{X}_{Oij}\boldsymbol{\beta}_O + \boldsymbol{\gamma}_{Oi}$ and covariance matrix $\boldsymbol{\rho}_O$ with the signs of latents depending on Y_{O1ij} and Y_{O2ij}, respectively, i.e. $Z_{Olij} > 0$ iff $Y_{Olij} = 1$. This can be implemented by iterating between draws from conditional truncated univariate normal distributions $f(Z_{O1ij}|Z_{O2ij})$ and $f(Z_{O2ij}|Z_{O1ij})$. For example, $[Z_{O1ij}|Z_{O2ij}]$ draws from the positive or negative part of a univariate normal distribution with mean $X_{O1ij}\beta_{O1} + \gamma_{O1i} + \rho_O(Z_{O2ij} - X_{O2ij}\beta_{O2} - \gamma_{O2i})$ and variance $1 - \rho_O^2$.

Step 2. Draw latent variables $\{Z_{R1ij}, Z_{R2ij}\}$. Similar to Step 1, Z_{R1ij} and Z_{R2ij} can be drawn jointly from a truncated bivariate normal distribution in SC_{11}, while $Z_{R1ij}(Z_{R2ij})$ is drawn from a truncated univariate normal distribution in $SC_{10}(SC_{01})$. Neither is drawn in SC_{00}.

Step 3. Assuming a flat prior for $\boldsymbol{\beta}_O$, draw it from $[\boldsymbol{\beta}_O|\text{others}]$

$$\sim N\Big(\big(\sum_{i,j\in S} \mathbf{X}_{Oij}^T \boldsymbol{\rho}_O^{-1} \mathbf{X}_{Oij}\big)^{-1} \big(\sum_{i,j\in S} \mathbf{X}_{Oij}^T \boldsymbol{\rho}_O^{-1} (Z_{Oij} - \boldsymbol{\gamma}_{Oi})\big), \big(\sum_{i,j\in S} \mathbf{X}_{Oij}^T \boldsymbol{\rho}_O^{-1} \mathbf{X}_{Oij}\big)^{-1}\Big).$$

Step 4. Assuming a flat prior, the posterior distribution of $\boldsymbol{\beta}_{(R)}$ is proportional to the product of the bivariate normal density of $\{Z_{R1ij}, Z_{R2ij}\}$ over SC_{11} and the two univariate normal densities of $\{Z_{R1ij}\}$ and $\{Z_{R2ij}\}$ over SC_{10} and SC_{01}, respectively. Applying the technique of combining multiple normals, the posterior distribution of $\boldsymbol{\beta}_R$ is shown to be

$$\boldsymbol{\beta}_R \sim N(\boldsymbol{\mu}_{\boldsymbol{\beta}_R}, \boldsymbol{\Omega}_{\boldsymbol{\beta}_R}),$$

where

$$\boldsymbol{\Omega}_{\boldsymbol{\beta}_R} = \left(\sum_{i,j \in SC_{11}} \mathbf{X}_{Rij}^T \boldsymbol{\rho}_R^{-1} \mathbf{X}_{Rij} + \begin{pmatrix} \sum_{i,j \in SC_{10}} X_{R1ij}^T X_{R1ij} & 0 \\ 0 & \sum_{i,j \in SC_{01}} X_{R2ij}^T X_{R2ij} \end{pmatrix} \right)^{-1},$$

and

$$\boldsymbol{\mu}_{\boldsymbol{\beta}_R} = \boldsymbol{\Omega}_{\boldsymbol{\beta}_R} \left(\sum_{i,j \in SC_{11}} \mathbf{X}_{Rij}^T \boldsymbol{\rho}_R^{-1} (Z_{Rij} - \gamma_{Ri} - \boldsymbol{\alpha}) + \begin{pmatrix} \sum_{i,j \in SC_{10}} X_{R1ij}^T (Z_{R1ij} - \gamma_{R1ij}) \\ \sum_{i,j \in SC_{01}} X_{R2ij}^T (Z_{R2ij} - \gamma_{R2ij}) \end{pmatrix} \right).$$

Step 5. Draw $\boldsymbol{\alpha}$ from

$$[\boldsymbol{\alpha}|\text{others}] \sim N\left(\sum_{i,j \in SC_{11}} \frac{(Z_{Rij} - \mathbf{X}_{Rij}\boldsymbol{\beta}_R - \boldsymbol{\gamma}_{Ri})}{n_{SC_{11}}}, \frac{\boldsymbol{\rho}_R}{n_{SC_{11}}} \right),$$

where $n_{SC_{11}}$ is the number of individuals in SC_{11} at S_1.

Step 6. Draw random effects $\boldsymbol{\gamma}_i$. The posterior density of $\boldsymbol{\gamma}_i$ is proportional to the product of two bivariate normal densities of $\{Z_{O1ij}, Z_{O2ij}\}$ and $\{Z_{R1ij}, Z_{R2ij}\}$, two univariate normal densities of $\{Z_{R1ij}\}$ and $\{Z_{R2ij}\}$, and the normal prior for $\boldsymbol{\gamma}_i$, and is a normal with covariance matrix

$$\boldsymbol{\Omega}_{\boldsymbol{\gamma}_i} = \left(\boldsymbol{\Sigma}^{-1} + \begin{pmatrix} n_i \boldsymbol{\rho}_O^{-1} & 0 \\ 0 & n_{i,SC_{11}} \boldsymbol{\rho}_R^{-1} + \begin{pmatrix} n_{i,SC_{10}} & 0 \\ 0 & n_{i,SC_{01}} \end{pmatrix} \end{pmatrix} \right)^{-1},$$

and mean

$$\boldsymbol{\mu}_{\boldsymbol{\gamma}_i} = \boldsymbol{\Omega}_{\boldsymbol{\gamma}_i}^{-1} \begin{pmatrix} \boldsymbol{\rho}_O^{-1} \sum_j (Z_{Oij} - \mathbf{X}_{Oij}\boldsymbol{\beta}_O) \\ \boldsymbol{\rho}_R^{-1} \sum_{j \in SC_{11}} (Z_{Rij} - \mathbf{X}_{Rij}\boldsymbol{\beta}_R - \boldsymbol{\alpha}) + \begin{pmatrix} \sum_{j \in SC_{10}} (Z_{R1ij} - X_{R1ij}\beta_{R1}) \\ \sum_{j \in SC_{01}} (Z_{R2ij} - X_{R2ij}\beta_{R2}) \end{pmatrix} \end{pmatrix},$$

where $n_{i,SC_{11}}$, $n_{i,SC_{10}}$, and $n_{i,SC_{01}}$ are the number of subjects in SC_{11}, SC_{10}, and SC_{01} at cluster i from S_1, respectively. If we simplify the model assumption and only include the random effects $\boldsymbol{\gamma}_{Oi}$, as in our application, they can be drawn from the normals with mean and covariance as the corresponding subparts of $\boldsymbol{\mu}_{\gamma_i}$ and $\boldsymbol{\Omega}_{\gamma_i}$, respectively.

Step 7. Draw ρ_O and ρ_R. The densities $[\rho_O|\text{others}]$ and $[\rho_R|\text{others}]$ are proportional to the bivariate normal densities of $\{Z_{O1ij}, Z_{O2ij}\}$ and $\{Z_{R1ij}, Z_{R2ij}\}$. Since neither of them has a closed form, they are sampled using the adaptive rejection Metropolis sampling.

Step 8. Draw $\boldsymbol{\Sigma}$ from

$$[\boldsymbol{\Sigma}|\text{others}] \sim IW(m+\nu, (\sum_{i \in S} \boldsymbol{\gamma}_i \boldsymbol{\gamma}_i^T + \Lambda^{-1})^{-1}).$$

Step 9. Imputation of $\{Y_{O1ij}^{mis}\}$ and $\{Y_{O2ij}^{mis}\}$ involves estimating conditional probabilities of $P(Y_{O1ij}, Y_{O2ij}|Y_{R1ij}, Y_{R2ij}, \text{others})$. When $Y_{Rlij} = 1$, Y_{Olij} is deterministically set to 1. When $Y_{Rlij} = 0$ for one of both of $l = 1, 2$, the corresponding probability or joint probabilities of $Y_{Olij} = 1$ can be calculated by straightforward application of Bayes rule using conditional probabilities that depend on univariate or bivariate normal cumulative distributions.

2 Tables

Table 2: Estimates from the bivariate model

Predictor	Chemotherapy				Radiation therapy			
	$\hat{\beta}_O$	$\widehat{\text{SE}}(\beta_O)$	$\hat{\beta}_R$	$\widehat{\text{SE}}(\beta_R)$	$\hat{\beta}_O$	$\widehat{\text{SE}}(\beta_O)$	$\hat{\beta}_R$	$\widehat{\text{SE}}(\beta_R)$
Intercept	0.097	0.128	1.195*	0.228	-1.204*	0.119	-0.027	0.525
Age								
65-74	REF							
18-54	0.280*	0.064	0.486*	0.140	0.245*	0.066	0.387	0.263
55-64	0.213*	0.061	0.345*	0.135	0.195*	0.062	0.356	0.243
75-84	-0.763*	0.054	0.134	0.136	-0.363*	0.064	-0.006	0.260
85+	-1.789*	0.100	-0.291	0.406	-0.911*	0.149	-1.035	0.643
Sex								
Male	REF							
Female	0.001	0.040	-0.104	0.100	-0.229*	0.044	0.246	0.186
Race								
White	REF							
Black	0.085	0.091	-0.201	0.223	-0.069	0.098	-0.348	0.414
Hispanic	0.103	0.070	-0.096	0.170	0.020	0.069	0.226	0.326
Asian	0.044	0.072	0.006	0.154	0.136	0.077	-0.428	0.275
Cancer type								
Stage 3 colon	REF							
Stage 2 rectal	-0.526*	0.057	0.440	0.236	1.532*	0.065	0.623*	0.228
Stage 3 rectal	0.205*	0.067	0.213	0.189	1.857*	0.062	0.596*	0.218
Income								
$>$ 50K	REF							
5-25K	-0.129*	0.060	-0.067	0.151	-0.083	0.063	0.132	0.275
30-35K	-0.054	0.062	-0.045	0.145	-0.045	0.061	0.069	0.264
40-50K	-0.083	0.051	0.174	0.132	0.039	0.058	-0.176	0.233
Marital status								
Unmarried	REF							
Married	0.289*	0.044	-0.206*	0.107	0.131*	0.046	-0.143	0.191
Comorbidity	-0.100*	0.014	-0.027	0.041	-0.090*	0.018	0.054	0.083
Hospital transfer								
No	REF							
Yes	0.178*	0.092	0.317	0.244	0.219*	0.096	0.376	0.412
Hospital volume								
High volume	REF							
Low volume	0.144	0.102	-0.590*	0.164	0.170	0.093	-0.649*	0.265
Medium volume	0.164	0.097	-0.548*	0.117	0.050	0.066	-0.038	0.219
ACOS hospital								
No	REF							
Yes	0.363*	0.077	0.031	0.107	0.166*	0.060	0.252	0.188
Rural hospital								
No	REF							
Yes	0.128	0.081	-0.290*	0.150	0.099	0.081	-0.332	0.290
Teaching hospital								
No	REF							
Yes	-0.207*	0.110	0.284	0.165	-0.128	0.086	0.555	0.383
Within survey region								
No	REF							
Yes	0.457*	0.072	NA	NA	0.107	0.059	NA	NA
Treated in 96 or 97								
No	REF							
Yes	0.098*	0.031	NA	NA	0.069*	0.035	NA	NA
Receipt of Radiation								
No	REF							
Yes	NA	NA	0.477*	0.144	NA	NA	NA	NA
Receipt of Chemo								
No	REF							
Yes	NA	NA	NA	NA	NA	NA	0.458	0.305
Correlation coefficient	$\hat{\rho}_O$	$\widehat{\text{SE}}(\rho_O)$	$\hat{\rho}_R$	$\widehat{\text{SE}}(\rho_R)$				
	0.704*	0.028	0.770*	0.074				
Random effects variance	$\hat{\sigma}^2_{O,chemo}$	$\widehat{\text{SE}}$	$\hat{\sigma}^2_{O,radiation}$	$\widehat{\text{SE}}$	$\widehat{\text{COR}}_O$	$\widehat{\text{SE}}$		
	0.238*	0.032	0.087*	0.015	0.104	0.105		

Note: * denotes that 95% credible interval does not contain 0.

Table 3: Full posterior predictive checking results

Testing Statistics	Survey Estimates	Univariate models 90% CI	Univariate models P-value	Bivariate model 90% CI	Bivariate model P-value
$\bar{Y}_{O1\cdot\cdot}$	0.733	(0.700, 0.741)	0.84	(0.695, 0.737)	0.91
$\bar{Y}_{O2\cdot\cdot}$	0.254	(0.238, 0.277)	0.45	(0.239, 0.278)	0.38
$\bar{Y}_{R1\cdot\cdot}$	0.614	(0.579, 0.628)	0.77	(0.577, 0.624)	0.84
$\bar{Y}_{R2\cdot\cdot}$	0.220	(0.199, 0.235)	0.63	(0.199, 0.236)	0.60
$P(Y_{R1ij} = 1 \vert Y_{O1ij} = 1)$	0.838	(0.811, 0.861)	0.55	(0.813, 0.862)	0.52
$P(Y_{R2ij} = 1 \vert Y_{O2ij} = 1)$	0.864	(0.802, 0.886)	0.78	(0.802, 0.886)	0.79
$\mathrm{OR}(Y_{O1ij}, Y_{O2ij})$	4.958	(1.168, 1.869)	1	(3.589, 7.167)	0.48
$\mathrm{OR}(Y_{O1ij}, Y_{R2ij})$	5.536	(1.186, 1.977)	1	(3.870, 8.395)	0.52
$\mathrm{OR}(Y_{O2ij}, Y_{R1ij})$	4.522	(1.315, 2.030)	1	(3.543, 6.514)	0.41
$\mathrm{OR}(Y_{R1ij}, Y_{R2ij})$	7.825	(1.378, 2.197)	1	(5.002, 10.073)	0.70
$\mathrm{VAR}(\bar{Y}_{O1i\cdot})$	0.0330	(0.0394, 0.0632)	0.003	(0.0407, 0.0636)	0.001
$\mathrm{VAR}(\bar{Y}_{O2i\cdot})$	0.0457	(0.0323, 0.0546)	0.65	(0.0327, 0.0558)	0.63
$\mathrm{COR}(\bar{Y}_{O1i\cdot}, \bar{Y}_{O2i\cdot})$	0.178	(-0.164, 0.282)	0.78	(0.0208, 0.415)	0.28
$\mathrm{VAR}(\bar{Y}_{R1i\cdot})$	0.0625	(0.0481, 0.0701)	0.71	(0.0486, 0.0695)	0.70
$\mathrm{VAR}(\bar{Y}_{R2i\cdot})$	0.0368	(0.0239, 0.0460)	0.63	(0.0241, 0.0467)	0.62
$\mathrm{COR}(\bar{Y}_{R1i\cdot}, \bar{Y}_{R2i\cdot})$	0.270	(-0.128, 0.321)	0.88	(0.106, 0.502)	0.31
$\mathrm{COR}(\bar{Y}_{O1i\cdot}, \bar{Y}_{R1i\cdot})$	0.507	(0.560, 0.833)	0.01	(0.582, 0.839)	0.009
$\mathrm{COR}(\bar{Y}_{O2i\cdot}, \bar{Y}_{R2i\cdot})$	0.728	(0.617, 0.946)	0.20	(0.611, 0.942)	0.23
$\mathrm{COR}(\bar{Y}_{O1i\cdot}, \bar{Y}_{R2i\cdot})$	0.104	(-0.182, 0.256)	0.62	(-0.012, 0.378)	0.18
$\mathrm{COR}(\bar{Y}_{O2i\cdot}, \bar{Y}_{R1i\cdot})$	0.160	(-0.141, 0.319)	0.67	(0.056, 0.486)	0.18

Note: The indexes i and j denote hospital and patients within the hospitals in the survey, respectively.

Table 4: Prediction of receiving radiation therapy for colon III cancer patients using logistic regression model

	Using the Survey		Using the Registry		Univariate model Imputation		Bivariate model Imputation	
Predictor	EST	SE	EST	SE	EST	SE	EST	SE
Age								
65-74	REF							
18-54	0.314	0.296	0.417*	0.115	0.321*	0.142	0.308*	0.128
55-64	0.402	0.272	0.437*	0.109	0.376*	0.139	0.334*	0.140
75-84	0.179	0.304	-0.168	0.127	-0.329*	0.149	-0.115	0.148
85+	0.207	0.681	-0.948*	0.373	-0.781*	0.366	-0.409	0.377
Chemotherapy								
No	REF							
Yes	1.700*	0.440	1.806*	0.124	1.079*	0.159	2.059*	0.269
Sex								
Male	REF							
Female	-0.318	0.211	-0.320*	0.085	-0.367*	0.108	-0.466*	0.096
Race								
White	REF							
Black	-0.082	0.475	-0.356	0.188	-0.072	0.221	-0.244	0.219
Hispanic	0.556	0.312	0.190	0.198	0.155	0.147	0.120	0.134
Asian	-0.664	0.369	0.157	0.148	0.177	0.170	0.228	0.182
Income								
$>$ 50K	REF							
5-25K	0.098	0.311	0.078	0.123	-0.045	0.151	0.067	0.164
30-35K	-0.047	0.293	0.090	0.120	0.025	0.138	0.090	0.138
40-50K	-0.071	0.275	-0.033	0.114	0.100	0.136	0.113	0.128
Marital status								
Unmarried	REF							
Married	0.078	0.228	-0.015	0.092	0.023	0.109	0.030	0.126
Comorbidity	-0.011	0.084	-0.056	0.038	-0.085	0.046	-0.089*	0.045
Hospital transfer								
No	REF							
Yes	0.297	0.433	0.240	0.156	0.236	0.208	0.322	0.215

Note: * denotes significance at 5% level.