Weighting Estimators for Cox Regression for studying etiological heterogeneity with partially observed multiple markers

Jooyoung Lee

Department of Applied Statistics

Chung-Ang University

2021 한국통계학회

With Shuji Ogino and Molin Wang

Etiological heterogeneity in Epidemiologic Research

- Molecular pathology investigates inherent individual heterogeneity of pathogenesis and disease processes (Ogino et al., 2013; Begg et al., 2013)

- The specific molecular subtypes of disease are often defined by *multiple markers*

- Interest lies in the heterogeneity effect of risk factors across disease subtypes

- Methods have been developed for statistical analysis (Chatterjee et al., 2010, Wang et al., 2016)

Missing Subtype data

- Missingness occurs due to the unavailability of tissue to be examined or even though tissue is available, markers' information can be missing
- Due to missing marker data, an event is known to occur, but the specific cause (or the specific subtype in our context) defined by markers is unknown
- The source of missing data may induce selection bias, such that the complete cases (i.e., cases with complete subtype data) do not represent the entire group of cases
- For multiple markers, some cases with subtype data unavailable may have partial information about the subtype due to missingness of some markers, not all the markers.

Nurses' Health Study (NHS) Data Set

id	time	cancer	agemo	period	smoke	tissue	**msi**	**braf**	**cimp**	**kras**
1	39	0	620	1	0	-	-	-	-	-
1	9	0	659	2	0	-	-	-	-	-
1	27	0	668	3	0	-	-	-	-	-
1	24	1	695	4	0	1	0	1	1	1

For **tumor markers**, 0 = missing, 1 = wild or high, 2 = mutant or low/negative

- Out of a total of 91,293 participants, 1482 colorectal cancer (CRC) cases
- 708 (47.8%) cases did not have tissue to be examined
- For 774 cases with available tissue, 610 subjects had complete tumor markers
- MSI: 104, CIMP: 106, BRAF: 104, KRAS: 149 missing turmor markers

Statistical Models

The cause-specific hazards Cox model

$$\begin{aligned}\lambda_{\mathbf{z}}(t|\mathbf{X},\mathbf{W}) &= \lim_{\Delta t \downarrow 0} \Delta^{-1} P(t \le T < t + \Delta t, \mathbf{Z} = \mathbf{z} | T \ge t, \mathbf{X}, \mathbf{W}) \\ &= \lambda_{0\mathbf{z}}(t) \exp(\boldsymbol{\beta}_{\mathbf{z}}^T \mathbf{X} + \boldsymbol{\eta}^T \mathbf{W})\end{aligned}$$

- $\tilde{T}$ be the time-to-event for the disease
- $T = \min(\tilde{T}, C)$ the observed time where C denote the censoring time
- $Y(s) = I(s \le T)$ the at-risk indicator process
- $N(s) = I(T \le s)$, $N_{\mathbf{z}}(s) = I(T \le s, \mathbf{Z} = \mathbf{z})$
- $\mathbf{Z} = (Z_1, \ldots, Z_K)$: K marker variables
- $\mathbf{X}$: P dimensional unconstrained variables including exposures
- $\mathbf{W}$: constrained variables

Statistical Models

Assume the proportional hazard assumption between the baseline hazard ratios

$$\lambda_{\mathbf{z}}(t|\mathbf{X}, \mathbf{W}) = \lambda_{0\mathbf{1}}(t) \exp(\alpha_{\mathbf{z}} + \boldsymbol{\beta}_{\mathbf{z}}^T \mathbf{X} + \boldsymbol{\eta}^T \mathbf{W})$$

We use log-linear models for the baseline hazard ratios and the covariate hazard ratios

Chatterjee et al. (2010)

$$\alpha_{\mathbf{z}} = \sum_{k=1}^{K} \xi_{k(z_k)}^{(1)} + \sum_{k=1}^{K} \sum_{k'>k}^{K} \xi_{kk'(z_k, z_{k'})}^{(2)} + \cdots + \xi_{12\cdots K(z_1,\ldots,z_K)}^{(K)}$$

$$\beta_{\mathbf{z}p} = \theta^{(0)p} + \sum_{k=1}^{K} \theta_{k(z_k)}^{(1)p} + \sum_{k=1}^{K} \sum_{k'>k}^{K} \theta_{kk'(z_k, z_{k'})}^{(2)p} + \cdots + \theta_{12\cdots K(z_1,\ldots,z_K)}^{(K)p}$$

- $\xi_{12\cdots k(z_1,\ldots,z_k)}^{(k)}$ the kth order parameter contrast for the log of cause-specific baseline hazard ratio
- $\theta^{(0)p}$ the regression coefficients for the reference disease subtype
- $\theta_{12\cdots k(z_1,\ldots,z_k)}^{(k)p}$ the kth order parameter contrasts for the log of hazard ratio for the covariate X_p.

Notation

- $\boldsymbol{\xi} = (\xi^{(1)}_{1(z_1)}, \ldots, \xi^{(2)}_{12(z_1,z_2)}, \ldots, \xi^{(K)}_{12\cdots K(z_1,\ldots,z_K)})$

- $\boldsymbol{\theta}_p = (\theta^{(0)p}, \theta^{(1)p}_{1(z_1)}, \ldots, \theta^{(2)p}_{12(z_1,z_2)}, \ldots, \theta^{(K)p}_{12\cdots K(z_1,\ldots,z_K)})$ for the pth element of $\boldsymbol{\theta}$

- $\boldsymbol{\beta}_{\mathbf{z}} = \boldsymbol{\theta}^T \boldsymbol{\mathcal{B}}_{\mathbf{z}}$ and $\alpha_{\mathbf{z}} = \boldsymbol{\xi}^T \mathcal{A}_{\mathbf{z}}$, where $\boldsymbol{\mathcal{B}}_{\mathbf{z}} = (\mathcal{B}_{\mathbf{z}1}, \ldots, \mathcal{B}_{\mathbf{z}P})$ with $\mathcal{B}_{\mathbf{z}p}$ the columns corresponding to marker $\mathbf{Z} = \mathbf{z}$ in the appropriate design matrix for covariate X_p, and similarly, we can define $\mathcal{A}_{\mathbf{z}}$

- $\bar{\mathbf{X}}_{\mathbf{Z}} = (\mathcal{A}_{\mathbf{Z}}, \mathcal{B}_{\mathbf{Z}1} \otimes \mathbf{X}_1, \ldots, \mathcal{B}_{\mathbf{Z}P} \otimes \mathbf{X}_P, \mathbf{W})$

- $\boldsymbol{\phi} = (\boldsymbol{\xi}, \boldsymbol{\theta}, \boldsymbol{\eta})$

Missing Subtype Data

- R_k be the missingness status of marker k (1=observed; 0=missing)
- $\mathbf{R} = (R_1, \ldots, R_K)$
- $\mathcal{D}_{\mathbf{R}}$ the set of all possible values of $\mathbf{R}$

The observed data are

$$\mathbf{O}_i = \{\delta_i, T_i, \delta_i \mathbf{R}_i, \delta_i \mathbf{R}_i \mathbf{Z}_i, \mathbf{X}_i, \mathbf{W}_i, \delta_i \mathbf{Q}_i\}$$

Missing-at-random assumption

$$\text{pr}(\mathbf{R}_i = \mathbf{r} | T_i, \delta_i = 1, \mathbf{V}_i, \mathbf{Z}_i) = \text{pr}(\mathbf{R}_i = \mathbf{r} | T_i, \delta_i = 1, \mathbf{V}_i),$$

where $\mathbf{V}_i = (T_i, \mathbf{X}_i, \mathbf{W}_i, \mathbf{Q}_i)$

Model for missingness $\mathbf{R}$

Conditional approach is given by Lipsitz and Ibrahim (1996)

$$
\begin{aligned}
&\mathrm{pr}(\mathbf{R}_i = \mathbf{r} | \delta_i = 1, \mathbf{V}_i; \boldsymbol{\psi}) \\
&\quad = \mathrm{pr}(R_{iK} = r_k | R_{i1} = r_1, \ldots, R_{i(K-1)} = r_{K-1}, \delta_i = 1, \mathbf{V}_i; \boldsymbol{\psi}_K) \times \cdots \\
&\quad \times \mathrm{pr}(R_{i1} = r_1 | \delta_i = 1, \mathbf{V}_i; \boldsymbol{\psi}_1)
\end{aligned}
$$

- $\pi_{\mathbf{r}}(\delta_i, \mathbf{V}_i; \boldsymbol{\psi}) = \mathrm{pr}(\mathbf{R}_i = \mathbf{r} | \delta_i = 1, \mathbf{V}_i; \boldsymbol{\psi}) I(\delta_i = 1) + I(\delta_i = 0)$
- Use logistic regression models

Two-stage missingness model

- R_{i0} indicate the tissue availability with 1 for being available and 0 for unavailable
- First, fit $\mathrm{pr}(R_{i0} = r_0 | \delta_i = 1, \mathbf{V}_i; \boldsymbol{\psi}_0)$
- Second, fit $\mathrm{pr}(\mathbf{R}_i = \mathbf{r} | \delta_i = 1, \mathbf{V}_i, R_{i0} = 1; \boldsymbol{\psi})$ for those with $R_{i0} = 1$

Estimating Equations

$\mathcal{U}_{\mathbf{Z}}$ denote the set of possible values of $\mathbf{Z}$

The approach of augmented data Lunn and McNeil (1995); Kalbfleisch and Prentice (2011)

id	time	cancer	agemo	period	smoke	tissue	**braf**	**cimp**	**kras**
1	**24**	**1**	**695**	**4**	**0**	**1**	**1**	**1**	**1**
1	24	0	695	4	0	1	1	1	2
1	24	0	695	4	0	1	1	2	1
1	24	0	695	4	0	1	1	2	2

$\vdots$

$$n^{-1}\sum_{i=1}^{n}\int_0^{\tau}\sum_{\mathbf{z}\in\mathcal{U}_{\mathbf{Z}}}\left\{\bar{\mathbf{X}}_{i\mathbf{z}}-\frac{\sum_{j=1}^{n}\sum_{\mathbf{z}\in\mathcal{U}_{\mathbf{Z}}}Y_j(t)\exp(\boldsymbol{\phi}^T\bar{\mathbf{X}}_{j\mathbf{z}})\bar{\mathbf{X}}_{j\mathbf{z}}}{\sum_{j=1}^{n}\sum_{\mathbf{z}\in\mathcal{U}_{\mathbf{Z}}}Y_j(t)\exp(\boldsymbol{\phi}^T\bar{\mathbf{X}}_{j\mathbf{z}})}\right\}dN_{i\mathbf{z}}(t)$$

Estimating Equations

Inverse Probability Weighted Estimator (IPW)

Horvitz and Thompson (1952); Gao and Tsiatis (2005)

$$U^{IPW}(\boldsymbol{\phi}, \hat{\boldsymbol{\psi}}) = n^{-1} \sum_{i=1}^{n} \int_0^{\tau} \sum_{\mathbf{z} \in \mathcal{U}_{\mathbf{Z}}} \frac{I(\mathbf{R}_i = \mathbf{1})}{\pi_{\mathbf{1}}(\delta_i, \mathbf{V}_i; \hat{\boldsymbol{\psi}})} \left\{ \bar{\mathbf{X}}_{i\mathbf{z}} - \frac{\tilde{\mathbf{S}}^{(1)}(\boldsymbol{\phi}, \hat{\boldsymbol{\psi}}, t)}{\tilde{S}^{(0)}(\boldsymbol{\phi}, \hat{\boldsymbol{\psi}}, t)} \right\} dN_{i\mathbf{z}}(t)$$

$$\tilde{\mathbf{S}}^{(a)}(\boldsymbol{\phi}; \boldsymbol{\psi}, t) = n^{-1} \sum_{i=1}^{n} \sum_{\mathbf{z} \in \mathcal{U}_{\mathbf{Z}}} \frac{I(\mathbf{R}_i = \mathbf{1})}{\pi_{\mathbf{1}}(\delta_i, \mathbf{V}_i; \boldsymbol{\psi})} Y_i(t) \exp(\boldsymbol{\phi}^T \bar{\mathbf{X}}_{i\mathbf{z}}) \bar{\mathbf{X}}_{i\mathbf{z}}^{\otimes a}, \quad a = 0, 1, 2$$

Augmented Inverse Probability Weighted Estimator (AIPW)

Robins et al. (1994); Gao and Tsiatis (2005)

$$U^{AIPW}(\boldsymbol{\phi}, \hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\gamma}}) = n^{-1} \sum_{i=1}^{n} \int_0^{\tau} \sum_{\mathbf{z} \in \mathcal{U}_{\mathbf{Z}}} \frac{I(\mathbf{R}_i = \mathbf{1})}{\pi_{\mathbf{1}}(\delta_i, \mathbf{V}_i; \boldsymbol{\psi})} \left\{ \bar{\mathbf{X}}_{i\mathbf{z}} - \frac{S^{(1)}(\boldsymbol{\phi}, t)}{S^{(0)}(\boldsymbol{\phi}, t)} \right\} dN_{i\mathbf{z}}(t)$$
$$- n^{-1} \sum_{i=1}^{n} D_i(\boldsymbol{\phi}, \hat{\boldsymbol{\psi}}_t, \hat{\boldsymbol{\gamma}})$$

$$D_i(\boldsymbol{\phi}; \hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\gamma}}) = \int_0^{\tau} \sum_{\mathbf{z} \in \mathcal{U}_{\mathbf{Z}}} \sum_{\mathbf{r} \neq \mathbf{1}} \left\{ \frac{I(\mathbf{R}_i = \mathbf{1})\pi_{\mathbf{r}}(\delta_i, \mathbf{V}_i; \hat{\boldsymbol{\psi}}) - I(\mathbf{R}_i = \mathbf{r})\pi_{\mathbf{1}}(\delta_i, \mathbf{V}_i; \hat{\boldsymbol{\psi}})}{\pi_{\mathbf{1}}(\delta_i, \mathbf{V}_i; \hat{\boldsymbol{\psi}})} \right\}$$
$$\times \left\{ \bar{\mathbf{X}}_{i\mathbf{z}} - \frac{\mathbf{S}^{(1)}(\boldsymbol{\phi}; t)}{S^{(0)}(\boldsymbol{\phi}; t)} \right\} \mathrm{pr}(\mathbf{Z}_i = \mathbf{z} | \delta_i = 1, \mathbf{V}_i, \mathbf{Z}_{i,obs_{\mathbf{r}}}; \hat{\boldsymbol{\gamma}}) dN_i(t)$$

$$\mathbf{S}^{(a)}(\boldsymbol{\phi}; t) = n^{-1} \sum_{i=1}^{n} \sum_{\mathbf{z} \in \mathcal{U}_{\mathbf{Z}}} Y_i(t) \exp(\boldsymbol{\phi}^T \bar{\mathbf{X}}_{i\mathbf{z}}) \bar{\mathbf{X}}_{i\mathbf{z}}^{\otimes a}, \quad a = 0, 1$$

- $\mathbf{Z}_{i,obs_r}$ be the observed components of marker $\mathbf{Z}_i$ when $\mathbf{R} = \mathbf{r}$

Model for markers $\mathbf{Z}$

Under the missing-at-random assumption, given $I(\delta_i = 1)$ and $\mathbf{V}_i$,

$$\mathrm{pr}(\mathbf{Z}_i | \delta_i = 1, \mathbf{V}_i; \boldsymbol{\gamma}) = \mathrm{pr}(\mathbf{Z}_i | \delta_i = 1, \mathbf{V}_i, \mathbf{R}_i = \mathbf{1}; \boldsymbol{\gamma})$$

We estimate $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)$ using a conditional logistic regression model

$$\mathrm{pr}(\mathbf{Z}_i = \mathbf{z} | \delta_i = 1, \mathbf{V}_i, \mathbf{R}_i = \mathbf{1}; \boldsymbol{\gamma}) = \frac{\exp(\boldsymbol{\gamma}_1^T \mathcal{A}_{\mathbf{z}} + \sum_{p=1}^{P} \boldsymbol{\gamma}_{2p}^T \boldsymbol{\mathcal{B}}_{\mathbf{z}p}^{-} \otimes X_{ip})}{\sum_{\upsilon \in \mathcal{U}_{\mathbf{Z}}} \exp(\boldsymbol{\gamma}_1^T \mathcal{A}_{\upsilon} + \sum_{p=1}^{P} \boldsymbol{\gamma}_{2p}^T \boldsymbol{\mathcal{B}}_{\upsilon p}^{-} \otimes X_{ip})}$$

The likelihood for the marker model is

$$\prod_{i=1}^{n} \prod_{\mathbf{z} \in \mathcal{U}_{\mathbf{Z}}} \left(\frac{\exp(\boldsymbol{\gamma}_1^T \mathcal{A}_{\mathbf{z}} + \sum_{p=1}^{P} \boldsymbol{\gamma}_{2p}^T \boldsymbol{\mathcal{B}}_{\mathbf{z}p}^{-} \otimes X_{ip})}{\sum_{\upsilon \in \mathcal{U}_{\mathbf{Z}}} \exp(\boldsymbol{\gamma}_1^T \mathcal{A}_{\upsilon} + \sum_{p=1}^{P} \boldsymbol{\gamma}_{2p}^T \boldsymbol{\mathcal{B}}_{\upsilon p}^{-} \otimes X_{ip})} \right)^{I(\mathbf{Z}_i = \mathbf{z}) I(\delta_i = 1, \mathbf{R}_i = 1)}$$

We denote $\rho_{\mathbf{z}}(\delta_i = 1, \mathbf{V}_i, \mathbf{R}_i = \mathbf{1}; \gamma) = \mathrm{pr}(\mathbf{Z}_i | \delta_i = 1, \mathbf{V}_i; \boldsymbol{\gamma})$

Simulation Studies

- Two markers which define four disease subtypes, denoted by (1,1), (1,2), (2,1), and (2,2)

- X unconstrained binary exposure with $\mathrm{pr}(X = 1) = 0.5$

- $\lambda_{\mathbf{z}}(t|X) = \lambda_{0\mathbf{1}}(t) \exp\left(\xi_{1(2)}^{(1)} + \xi_{2(2)}^{(1)} + \left\{\theta^{(0)} + \theta_{1(2)}^{(1)} + \theta_{2(2)}^{(1)}\right\} X\right)$

- For identifiability $\xi_{1(1)}^{(1)} = \xi_{2(1)}^{(1)} = \theta_{1(1)}^{(1)} = \theta_{2(1)}^{(1)} = 0$

- Weibull baselines with $\lambda_{0\mathbf{1}}(t) = \nu \lambda_{0\mathbf{1}} t^{\nu-1}$

- Censoring: $N(75, 5^2)$

- $\mathbf{R}$ depends on X and T

- Sample size = 10,000 and Simulation replicates = 1,000

Table 1: Simulation results for one-stage proposed model with two markers, each with two levels. In Case 1 both $\pi_{\mathbf{r}}(\cdot)$ and $\rho_{\mathbf{z}}(\cdot)$ were correctly specified with sample size of 10000 and 1000 simulation replicates.

Approach	$\theta_0^{(0)}$ (truth: 0.35)				$\theta_{1(2)}^{(1)}$ (truth: 0.00)				$\theta_{2(2)}^{(1)}$ (truth: 0.25)			
	% BIAS	ESE	ASE	CP	BIAS	ESE	ASE	CP	% BIAS	ESE	ASE	CP
					z_1 : 50%; z_2 : 45% *missing*							
Full	-0.005	0.078	0.075	0.946	0.002	0.107	0.104	0.944	0.020	0.169	0.166	0.941
CCA	-2.302	0.140	0.138	0.000	0.004	0.193	0.190	0.952	-0.028	0.294	0.301	0.951
EE	-0.042	0.160	0.139	0.940	0.001	0.209	0.214	0.952	0.499	0.343	0.412	0.971
(CASE 1)												
IPW	-0.010	0.127	0.128	0.954	0.003	0.200	0.199	0.950	-0.036	0.300	0.310	0.964
AIPW	-0.005	0.104	0.102	0.946	0.001	0.170	0.167	0.948	-0.026	0.241	0.250	0.961

Table 1 continued. In Case 2 $\pi_{\mathbf{r}}(\cdot)$ was correctly specified but $\rho_{\mathbf{z}}(\cdot)$ was misspecified, in Case 3 $\rho_{\mathbf{z}}(\cdot)$ was correctly specified but $\pi_{\mathbf{r}}(\cdot)$ was misspecified, and in Case 4 both $\pi_{\mathbf{r}}(\cdot)$ and $\rho_{\mathbf{z}}(\cdot)$ were misspecified with sample size of 10000 and 1000 simulation replicates.

Approach	$\theta_0^{(0)}$ (truth: 0.35)				$\theta_{1(2)}^{(1)}$ (truth: 0.00)				$\theta_{2(2)}^{(1)}$ (truth: 0.25)			
	% BIAS	ESE	ASE	CP	BIAS	ESE	ASE	CP	% BIAS	ESE	ASE	CP
(CASE 2)												
IPW	-0.010	0.127	0.128	0.954	0.003	0.200	0.199	0.950	-0.036	0.300	0.310	0.964
AIPW	-0.006	0.104	0.102	0.946	0.001	0.170	0.167	0.948	-0.025	0.241	0.250	0.960
(CASE 3)												
IPW	-0.031	0.121	0.124	0.957	0.004	0.193	0.193	0.957	-0.028	0.294	0.301	0.952
AIPW	-0.002	0.101	0.099	0.942	0.002	0.165	0.162	0.950	-0.005	0.236	0.243	0.953
(CASE 4)												
IPW	-0.031	0.121	0.124	0.957	0.004	0.193	0.193	0.957	-0.028	0.294	0.301	0.952
AIPW	-0.007	0.101	0.099	0.943	0.002	0.165	0.162	0.950	-0.018	0.236	0.243	0.954

Application to NHS study

- Exposure: pack-years of smoking before age of 30 (no, <5 pack-years, 5 $\geq$ pack-years)
- 4 binary tumor markers: MSI (high vs. MSS), CIMP (high vs. low/negative), BRAF (wild vs. mutant), KRAS (mutation vs. mutant)
- 16 possible colorectal cancer subtypes
- Variables adjusted for: body mass index (kg/m2, continuous), regular aspirin use (yes or no), family history of CRC (yes or no), alcohol intake (0.0-0.14, 0.15-1.9, 2.0-7.4, $\geq$ 7.5 g/day), physical activity (<5, 5-11.4, 11.5-21.9, $\geq$22 MET-hours/week)
- Variables for the missingness model
  - Logistic regression model for the first stage: age at CRC (months) + tumor location (proximal = 1, distal = 2, rectum = 3, others = 4) Colussi et al. (2013)
  - Multinomial logistic regression model for the second stage: age at CRC (months) + tumor location (proximal = 1, distal = 2, rectum = 3, others = 4)

Table 2: Results of the NHS (1986-2012) data analysis for modeling the pack-years of smoking before age of 30 and CRC subtype association using 4 binary markers: MSI, CIMP, *BRAF*, and *KRAS*

				MSI		CIMP		*BRAF*		*KRAS*	
Method		$\theta^{(0)1}$	$\theta^{(0)2}$	$\theta^{(1)1}_{1(2)}$	$\theta^{(1)2}_{1(2)}$	$\theta^{(1)1}_{2(2)}$	$\theta^{(1)2}_{2(2)}$	$\theta^{(1)1}_{3(2)}$	$\theta^{(2)2}_{3(2)}$	$\theta^{(1)1}_{4(2)}$	$\theta^{(1)2}_{4(2)}$
CCA	EST	0.122	-0.189	0.165	0.569	-0.019	-0.003	-0.133	0.298	0.020	0.368
	SE	0.243	0.232	0.354	0.314	0.231	0.197	0.353	0.309	0.312	0.289
	p-value	0.614	0.417	0.641	0.070	0.936	0.989	0.707	0.334	0.949	0.202
IPW	EST	0.117	-0.184	0.162	0.551	-0.054	-0.021	-0.134	0.297	0.068	0.344
	SE	0.237	0.225	0.354	0.311	0.236	0.201	0.349	0.304	0.309	0.283
	p-value	0.622	0.415	0.646	0.076	0.819	0.918	0.702	0.328	0.826	0.225
AIPW	EST	-0.109	-0.221	0.233	**0.647**	-0.041	-0.007	-0.049	0.374	0.096	0.402
	SE	0.225	0.216	0.342	**0.305**	0.222	0.193	0.338	0.300	0.316	0.291
	p-value	0.629	0.306	0.496	**0.034**	0.852	0.970	0.885	0.213	0.762	0.166

Discussion

- To elucidate inherent heterogeneity of pathogenesis and disease processes among individuals, cancer subtypes are classified by multiple markers
- Appropriately address the selection bias by accounting for missingness explained by auxiliary variables
- Make use of all available data, not only complete-cases
- Provide protection against the misspecification of either missingness models or marker models due to the double-robustness property
- Our proposed AIPW method can provide efficient and valid estimation exploiting all available data in the era of molecular pathological epidemiology.

References

OGINO, S., LOCHHEAD, P., CHAN, A. T., NISHIHARA, R., CHO, E., WOLPIN, B. M., MEYERHARDT, J. A.,MEISSNER, A., SCHERNHAMMER, E. S., FUCHS, C. S. et al. (2013). Molecular pathological epidemiology of epigenetics: emerging integrative science to analyze environment, host, and disease.Modern Pathology 26, 465.

BEGG, C. B., ZABOR, E. C., BERNSTEIN, J. L., BERNSTEIN, L., PRESS, M. F. & SESHAN, V. E. (2013). A conceptual and methodological framework for investigating etiologic heterogeneity. Statistics in medicine 32, 4955039–5052.

CHATTERJEE, N., SINHA, S., DIVER, W. R. & FEIGELSON, H. S. (2010). Analysis of cohort studies with multivariate and partially observed disease classification data. Biometrika 97, 683–698.

WANG, M., SPIEGELMAN, D., KUCHIBA, A., LOCHHEAD, P., KIM, S., CHAN, A. T., POOLE, E. M., TAMIMI,R., TWOROGER, S. S., GIOVANNUCCI, E., BERNARD, R. & OGINO, S. (2016). Statistical methods for studying disease subtype heterogeneity. Statistics in Medicine 35, 782–800.

LIU, L., NEVO, D., NISHIHARA, R., CAO, Y., SONG, M., TWOMBLY, T., CHAN, A., GIOVANNUCCCI, E., VANDERWEELE, T., WANG, M. & OGINO, S. (2017). Utility of inverse probability weighting in molecular pathologcal epidemiology. European Journal of Epidemiology.

LIPSITZ, S. R. & IBRAHIM, J. G. (1996). A conditional model for incomplete covariates in parametric regression models. Biometrika 83, 916–922.

LUNN, M. & MCNEIL, D. (1995). Applying cox regression to competing risks. Biometrics 51, 524–532.

KALBFLEISCH, J. D. & PRENTICE, R. L. (2011). The Statistical Analysis of Failure Time Data, vol. 360. John Wiley & Sons.

HORVITZ, D. G. & THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 47, 663–685.

GAO, G. & TSIATIS, A. A. (2005). Semiparametric estimators for the regression coefficients in the linear transfor- mation competing risks model with missing cause of failure. Biometrika 92, 875–891.

ROBINS, J. M., ROTNITZKY, A. & ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association 89, 846–866.

COLUSSI, D., BRANDI, G., BAZZOLI, F. & RICCIARDIELLO, L. (2013). Molecular pathways involved in colorectal cancer: implications for disease behavior and prevention. International journal of molecular sciences 14, 16365– 16385.

NEVO, D., NISHIHARA, R., OGINO, S. & WANG, M. (2018). The competing risks cox model with auxiliary case covariates under weaker missing-at-random cause of failure. Lifetime data analysis 24, 425–442.