Statistics and Its Interface Volume 12 (2019) 423–437

# Adjusted-crude-incidence analysis of multiple treatments and unbalanced samples on competing risks

Sangbum Choi, Chaewon Kim, Hua Zhong, Eun-Seok Ryu*, and Sung Won Han*

In this paper, we discuss adjusted cumulative incidence in multiple treatment groups with unbalanced samples. In a nonrandomized experiment or an observational study, the observed data may be unbalanced in covariates when multiple treatments are administered differently based on patients' characteristics. In the case of multiple survival outcomes, clinical researchers are often interested in estimating the cumulative incidence within a specific treatment group, and this approach is subject to a potential bias with unbalanced samples. Using extensive simulation analyses, we demonstrate that a naïve approach to the estimation of a cumulative incidence curve may yield misleading results, unless patients' characteristics are fully considered. To achieve an unbiased estimation from unbalanced data, we propose an adjusted cumulative incidence based on the inverse probability of a treatment weighting. In a series of simulations, the proposed method shows robust performance when estimating cumulative incidence under various scenarios, including balanced and unbalanced samples. Lastly, we explain how to apply the proposed method using an example based on real data.

Keywords and phrases: Competing risks, Cumulative incidence, Inverse probability of treatment weighting, Kaplan–Meier, Survival analysis.

## 1. INTRODUCTION

Data on competing risks are commonly obtained in biomedical research, particularly in cancer studies, where the need to deal with multiple potential outcomes is nearly ubiquitous; see Satagopan et al. (2004); Kim (2007); Dignam and Kocherginsky (2008); Lau et al. (2009); Andersen et al. (2012) for a comprehensive review of this topic. If there are competing risks, individuals are observed from when they enter the study until the occurrence of an event of interest, a competing event, or censoring. In oncology, for example, competing risks are encountered when cancer patients are followed after treatment, and their first failure event may be a local recurrence, distant metastasis, the onset of a second primary cancer, or death. A patient can potentially experience failure from multiple causes, but we might observe only the first event type, which possibly precludes other types of events. Even in cases where individuals can have subsequent events (e.g., cancer recurrence, followed by death), clinicians might focus on the occurrence of the first event to estimate disease-free survival (DFS), thus creating a problem of competing risks.

In analyses of competing risks, two principal factors—the cause-specific hazard (CSH) function and the cumulative incidence function (CIF)—are identifiable and, thus, are commonly used to summarize outcomes by event type. With multiple causes of failure, we would like to know the marginal or net probability of a failure caused by a given event type, in the absence of other failure types. Nonetheless, this is a hypothetical measure that typically involves untestable parametric assumptions about the dependence of the censoring mechanism on the underlying survival times (Kalbfleisch and Prentice, 2002, Chapter 8). The CSH quantifies the rate of transition to the event of interest in a real situation, in which individuals may instead experience another type of event. Alternatively, the CIF allows us to estimate the crude incidence of an event while taking competing risks into account. In the presence of competing risks, there is no longer a one-to-one correspondence between the CSH and the CIF for a specific event type. With covariates, several modeling approaches are available for evaluating the relation of covariates to cause-specific failures through the CSH or CIF (e.g., Prentice et al., 1978; Fine and Gray, 1999; Peng and Fine, 2009; Choi and Huang, 2014).

Although much has been written on estimations in situations of competing risks, less attention has been devoted to hypothesis testing related to multiple treatments with unbalanced samples. In nonrandomized experiments or observational data, samples are often unbalanced because multiple treatments are sometimes applied differently based on patients' characteristics, where a risk measure within a specific treatment group can be biased owing to the unbalanced samples. To overcome such problems, several established methods are used for stratification or matching in order to adjust the survival estimation (Hankey and Myers, 1971;

*Corresponding Authors.

Cupples et al., 1995; Amato, 1988; Nieto and Coresh, 1996). To assess overall survival rates, while accounting for potential confounders, Xie and Liu (2005) derived an appropriate weighted survival estimate, the so-called adjusted Kaplan–Meier estimator, which is based on the inverse probability weighting scheme of a treatment (Robins and Rotnitzky, 1992). The inverse probability of the treatment weighting (IPTW), which reflects the extent to which a patient belongs to the treatment group, given the covariates, is assigned to each observation, creating a potentially balanced pseudo-population. Nevertheless, no estimator of cumulative incidence has been proposed that accounts for competing events. In the presence of competing risks, the choice of test may be unclear, and substantively different inferential conclusions may arise from the same data, depending on which risk measure is used. Such discrepancies occur naturally, because the tests address different parameters of the failure process, one or more of which may be of interest in a given situation.

In this article, we propose a method for estimating a nonparametric cumulative (crude) incidence probability by means of multiple treatment groups with unbalanced samples. We first adjust the survival function by the IPTW, which serves as an estimate of the ordinary survival function of Xie and Liu (2005), and then use the adjusted survival function for the calculation of the crude risk probability. The proposed method shows good performance in simulations and is applied to the Glioma data set from The Cancer Genome Atlas (TCGA) project. The remainder of the paper is organized as follows. Section 2 presents the proposed weighted crude incidence estimator for unbalanced treatment data, along with background information on analyses of competing risks. Section 3 describes the simulation results on the performance of the proposed method on a finite sample. Section 4 applies the proposed method to a real-data example. In Section 5, we provide concluding remarks.

## 2. METHODS

### 2.1 Notation and basics

Let $T$ be the failure time variable, and suppose there are $K$ possible causes of failure, denoted by $\varepsilon = 1, 2, ..., K$. For simplicity, we assume there are two competing events (i.e., $K = 2$) because the event of interest is singled out and coded as 1, and all types of events other than a cause-1 event can be grouped together as cause-2 events. In the presence of competing risks, the CSH function of the $k$th event is defined as

$$\lambda_k(t) = \lim_{dt \to 0} \frac{1}{dt} P(t \le T < t + dt, \varepsilon = k | T \ge t),$$

and $\Lambda_k(t) = \int_0^t \lambda_k(s) ds$. The CSH function describes the instantaneous risk of event $k$ for subjects that are currently event-free. In the presence of covariates, researchers often adapted a marginal proportional hazards model (Prentice et al., 1978) to $\lambda_k(\cdot)$, in which prognostic factors associated with the biological mechanism behind event $k$ may change the instantaneous event risk. On the other hand, the probability that an event occurs in a specific period depends on the CSHs of the other events (Gray, 1988; Fine and Gray, 1999). The crude probability of an event type can be determined by the CIF, given by

$$F_k(t) = P(T \le t, \varepsilon = k).$$

Then, we can write $S(t) = P(T > t) = 1 - \sum_k F_k(t)$ to denote the overall probability of surviving any cause of failure.

A convenient representation of the CIF as a product limit estimator naturally arises starting from the subdistribution hazard introduced by Gray (1988), and has the form

$$\lambda_k^*(t) = \lim_{dt \to 0} \frac{1}{dt} P(t \le T < t + dt, \varepsilon = k | (T \ge t) \cup (T < t, \varepsilon \ne k)).$$

This hazard formulation has been shown to be useful when comparing the crude incidence between different groups because it restores the one-to-one relation between the subdistribution hazard and the cumulative probability of a particular failure type; that is,

$$\lambda_k^*(t) = -d \log\{1 - F_k(t)\}/dt.$$

One can think of $\lambda_k^*$ as the hazard function for an improper random variable $T_k^* = I(\varepsilon = k) \times \tilde{T} + I(\varepsilon \ne k) \times \infty$, which has a distribution function equal to $F_k(t), t < \infty$ and point mass $1 - F_k(\infty)$ at $t = \infty$. Note that for any finite $t$, $T_k^* \le t$ is equivalent to $T \le t$ and $\varepsilon = k$, implying that $P(T_k^* \le t) = F_k(t)$. The definition of $T_k^*$ is consistent with the argument that when an event other than $k$ occurs first, the latter will never be observed as having occurred first and, thus, the corresponding time will be infinity.

### 2.2 Adjusted crude incidence function (ACIF)

Let $(T_i, C_i, \varepsilon_i, X_i, \mathbf{Z}_i)$, $i = 1, 2, ..., n$, denote an independent sample of competing risk data with two or more groups, where $T_i$ and $C_i$ are the failure and censoring time variables, respectively, $X_i$ is the group index, $X_i = 1, 2, ..., R$ for $R$ different treatment groups, and $\mathbf{Z}_i$ is the $p$-dimensional covariate vector. In the presence of right censoring, the actual event time and the event type are partially observed, in which we can have $\tilde{T}_i = \min(T_i, C_i)$ and $\Delta_i = \delta_i \varepsilon_i$, the observed failure time and failure status, respectively, with $\delta_i = I(T_i \le C_i)$ denoting the censoring indicator. The observed competing risks data can be expressed as $(\tilde{T}_i, \Delta_i, X_i, \mathbf{Z}_i)$, which is an independent copy of $(\tilde{T}, \Delta, X, \mathbf{Z})$ for subject $i = 1, 2, ..., n$.

In this section, we discuss how to derive the overall and adjusted cumulative incidence functions. First, designate

the covariate index as $d \in \{1, ..., p\}$. For the sample index, $i = 1, 2, ..., n$, suppose $\pi_{ir}$ is the probability of the $i$th subject's being in group $r \in \{1, ..., R\}$. This probability may depend on the covariate vector $\mathbf{Z}_i$, (i.e., $\pi_{ir} = P(X_i = r|\mathbf{Z}_i)$). Here, we assume that (i) $\{\pi_{ir}\}$ are either known, as in a designed study, or can be consistently estimated given $\mathbf{Z}_i$, and (ii) $\{\pi_{ir}\}$ are bounded below by zero. Similarly, let $w_{ir}$ represent the weight for the $i$th subject who is assigned to the treatment group with $X_i = r$. When $\{\pi_{ir}\}$ are unknown and need to be estimated from the data, we may use a nonparametric smoothing method (Wang et al., 1997) or a parametric multinomial logistic regression that relates the probability $\pi_{ir}$ of assigning treatment $r$ to subject $i$, given the covariates, as

$$\begin{aligned}\pi_{ir} &\equiv P(X_i = r|\mathbf{Z}_i = \mathbf{z}) \\ &= \frac{\exp(\zeta_r + \gamma_r'\mathbf{z})}{\sum_{l=1}^{R} \exp(\zeta_l + \gamma_l'\mathbf{z})}, \quad r = 1, ..., R-1, \ i = 1, ..., n,\end{aligned}$$

where $\zeta_r$ is the intercept for the $r$th treatment, and $\gamma_r$ is a vector of the parameters associated with $\mathbf{z}$. For model identification, we set $(\zeta_R, \gamma_R) = (0, 0)$, and, thus $\pi_{iR} = 1 - \sum_{r=1}^{R-1} \pi_{ir}$ is fixed. Once the set of parameters $(\zeta_r, \gamma_r)$ is estimated, we may define a weight as $w_{ir} = I(X_i = r)/\pi_{ir}$, following the IPTW scheme. It is easy to see that $E(w_{ir}) = 1$ if the model for $\pi_{ir}$ is specified correctly.

Suppose that survival time values are observed at times $t_1 < t_2 < ... < t_D$. At time $t_j, j = 1, ..., D$, there are $(d_{1,jr}, d_{2,jr})$ individuals, representing cause-1 and cause-2 events, respectively, out of $Y_{jr}$ individuals at risk in group $r$. We can write $d_{k,jr} = \sum_{i:\tilde{T}_i = t_j} I(X_i = r, \Delta_i = k)$, $k = 1, 2$, and $Y_{jr} = \sum_{i:\tilde{T}_i \geq t_j} I(X_i = r)$, where the indicator function $I(X_i = r)$ is equal to one if the $i$th subject receives treatment $r$, and zero otherwise. Here, $d_{k,jr}$ indicates the number of patients who die of cause $k$ at time $t_j$ in group $r$. Then, $Y_{jr}$ indicates the number of patients who are still at risk until time $t_j$ in group $r$; therefore, they are at risk.

Accordingly, the weighted number of cause-specific events and the weighted number at risk in group $r$ are defined as

$$\begin{aligned}d_{k,jr}^w &= \sum_{i:\tilde{T}_i = t_j} w_{ir} I(X_j = r, \Delta_i = k) \\ &= \sum_{i:\tilde{T}_i = t_j} \frac{I(X_j = r, \Delta_i = k)}{\pi_{ir}}\end{aligned} \tag{1}$$

for $k = 1, 2$, and

$$Y_{jr}^w = \sum_{i:\tilde{T}_i \geq t_j} w_{ir} I(X_i = r) = \sum_{i:\tilde{T}_i \geq t_j} \frac{I(X_i = r)}{\pi_{ir}}, \tag{2}$$

respectively.

Several research groups, including Gail and Pfeiffer (2005) and Benichou and Gail (1990), have studied parametric crude incidence estimations. On the other hand, Klein and Moeschberger (2003) discuss the nonparametric estimation of $F_k(t)$, given by

$$\hat{F}_k(t) = \begin{cases} 0, & \text{if } t \leq t_1, \\ \sum_{j:t_j \leq t} \hat{S}(t_j-) \frac{d_{k,j}}{Y_j}, & \text{if } t_1 < t, \end{cases}$$

where $\hat{S}(t_j-)$ is the Kaplan–Meier estimator, evaluated immediately before time $t$, and $d_{k,j} = \sum_{r=1}^{R} d_{k,jr}$ and $Y_j = \sum_{r=1}^{R} Y_{jr}$ represent the number of cause-specific events and the number of individuals at risk, respectively, at time $t_j$. The CIF estimators and the direct testing procedures are available in the R packages, `cmprsk` and `timereg`, respectively.

To reduce the sample bias of different groups with competing risks, we propose estimating the ACIF for group $r$ using

$$\hat{F}_{kr}^w(t) = \sum_{j:t_j \leq t} \hat{S}_r^w(t_j-) \frac{d_{k,jr}^w}{Y_{jr}^w}, \tag{3}$$

where the adjusted estimate of the treatment-specific survival function, $\hat{S}_r^w(t)$, can be obtained as

$$\hat{S}_r^w(t) = \begin{cases} 1, & \text{if } t < t_1, \\ \prod_{j;t_j \leq t} \left[1 - \frac{d_{1,jr}^w + d_{2,jr}^w}{Y_{jr}^w}\right], & \text{if } t_1 \leq t. \end{cases} \tag{4}$$

If subjects are assigned equally to each treatment group, as in randomized clinical trials, the ACIF estimator $\hat{F}_{k,r}^w(t)$ is reduced to the conventional CIF estimator, $\hat{F}_k(t)$.

## 2.3 Variance estimation for the ACIF

For ease of exposition, here and below, we consider a case with only two treatment groups, denoted by $X_i = 0$ and 1. We refer to groups 1 and 0 as the treatment group and the control group, respectively, and let $\pi_i = P(X_i = 1|\mathbf{Z}_i)$ and $1 - \pi_i = P(X_i = 0|\mathbf{Z}_i)$. Then, the weight for the $i$th subject is $w_i = 1/\pi_i$ if she received treatment 1, and $w_i = 1/(1-\pi_1)$ otherwise. To derive the variance estimation of the ACIF, we use some of the intermediate results from Xie and Liu (2005), and assume that

$$\max_{i:\tilde{T}_i \geq t_j} \frac{(1/\pi_i)}{\sum_{i:\tilde{T}_i \geq t_j} (1/\pi_i)} \to 0 \ \text{ as } \ n \to \infty. \tag{5}$$

Xie and Liu (2005) demonstrated that $E[\hat{S}_1^w(t)] = S_1(t)$ if $t \leq T_{\max}$, where $T_{\max}$ denotes the largest observed failure time in the treatment group. Moreover, it turns out that under condition (5), the variance estimator of $\hat{S}_1^w(t)$ has the following formulation if either $\pi_i$ is known in advance or if $\pi_i$ is unknown and $\hat{\pi}_i$ need to be estimated from the data:

$$\widehat{\text{var}}[\hat{S}_1^w(t)] = [\hat{S}_1^w(t)]^2 \sum_{j:t_j \leq t} \frac{1 - \hat{\theta}_{j1}^w}{M_j \hat{\theta}_{j1}^w}, \tag{6}$$

where $\hat{\theta}^w_{j1} = 1 - \hat{\lambda}^w_{1,j1} - \hat{\lambda}^w_{2,j1}$, with $\hat{\lambda}^w_{k,jr} = d^w_{k,jr}/Y^w_{jr}$, $(k = 1,2, r = 1,2)$, and $M_j = \frac{(\sum_{i:T_i \geq t_j} 1/\pi_i)^2}{\sum_{i:T_i \geq t_j} (1/\pi_i)^2}$. Similar results hold for the control group at $X_i = 0$.

Using an argument similar to that of Xie and Liu (2005), we can show that, for instance, $E[\hat{F}^w_{11}(t)] = F_{11}(t)$ if $t \leq T_{\max}$, where $F_{11}(t)$ represents the CIF of the treatment group who failed as a result of cause 1. The asymptotic variance of $\hat{F}_{11}(t)$ can be obtained by applying the delta approximation, and its estimate for the interval $t_j \leq t < t_{j+1}$ is expressed as

$$
\begin{aligned}
(7)\quad & \widehat{\mathrm{var}}[\hat{F}^w_{11}(t)] \\
&= \sum_{l=1}^{j} \{\hat{\lambda}^w_{1,l1} \hat{S}^w_1(t_{l-1})\}^2 \left\{ \frac{1 - \hat{\lambda}^w_{1,l1}}{M_l \hat{\lambda}^w_{1,l1}} + \sum_{m=1}^{l-1} \frac{1 - \hat{\theta}^w_{m1}}{M_m \hat{\theta}^w_{m1}} \right\} \\
&\quad + 2 \sum_{l=1}^{j-1} \sum_{s=l+1}^{j} \{\hat{\lambda}^w_{1,l1} \hat{S}^w_1(t_{l-1}) \hat{\lambda}^w_{1,s1} \hat{S}^w_1(t_{s-1})\} \\
&\quad \times \left\{ -\frac{1}{Y^w_{l1}} + \sum_{m=1}^{l-1} \frac{1 - \hat{\theta}^w_{m1}}{M_m \hat{\theta}^w_{m1}} \right\}.
\end{aligned}
$$

The above asymptotic standard deviations of the ACIF can be applied to estimations of pointwise confidence intervals in analyses of competing risk data with group stratification. A linear $100(1-\alpha)\%$ pointwise confidence interval for $\hat{F}^w_{11}(t)$ for fixed time $t$ is

$$
(8)\qquad \hat{F}^w_{11}(t) + z_{\alpha/2} \hat{\sigma}_{11}(t),
$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution, and $\hat{\sigma}_{11}(t)$ is the square root of $\widehat{\mathrm{var}}[\hat{F}^w_{11}(t)]$, as given by (7). Because this interval might not be completely satisfactory for small sample sizes, we can also take advantage of the log-log transformation approach (Choudhury, 2002), which corresponds to the $100(1-\alpha)\%$ confidence interval of the form

$$
\exp\left\{ \frac{\pm z_{\alpha/2} \hat{\sigma}_{11}(t)}{\hat{F}^w_{11}(t) \log\{\hat{F}^w_{11}(t)\}} \right\}.
$$

## 3. SIMULATION ANALYSIS

### 3.1 An adjusted Kaplan–Meier curve for data from a proportional hazards model

The simulation structure is inspired by the structure specified by Xie and Liu (2005), who successfully compared the performance of a Kaplan–Meier curve and an adjusted Kaplan–Meier curve in terms of adjusting for the bias in unbalanced data. Here, we show how well the adjusted CIF can estimate the cumulative incidence curve, adjusting for the bias caused by unbalanced data.

A thousand samples are generated based on the covariates, $z_1$, $z_2$, $z_3$, and $z_4$. The covariate $z_1$ is simulated by means of a Bernoulli distribution, with $z_1 \approx$ Bernoulli(0.5), which indicates gender, and the covariate $z_2$ is simulated by means of a uniform distribution, with $z_2 \approx$ uniform(40,60), which indicates age. The covariate $z_3$ is generated from a uniform distribution, with $z_3 \approx$ uniform(40,60), which indicates weight, and the covariate $z_4$ is generated by sampling to return 0, 1, or 2, indicating certain categorical variables.

Based on the above setting, a group indicator variable $x_i$ is generated with the following probability.

$$
P(x = 1|z_1) = \begin{cases} 0.5 & \text{if } z_1 = 0 \\ 0.5 & \text{if } z_1 = 1, \end{cases}
$$

which means that we generate a balanced data set. We also generate an unbalanced data set for another scenario, where a group indicator variable $x_i$ is generated with probability

$$
P(x = 1|z_1) = \begin{cases} 0.8 & \text{if } z_1 = 1 \\ 0.2 & \text{if } z_1 = 0, \end{cases}
$$

which has the same conditions as the balanced data set except for the probability of the group indicator $x$.

In the simulation, we provide a different time mean for the group and covariate $z_1$' as follows:

(1) $\frac{1}{\lambda_{00}} = E(T|X = 0, z_1 = 0) = 20$
(2) $\frac{1}{\lambda_{01}} = E(T|X = 0, z_1 = 1) = 60$
(3) $\frac{1}{\lambda_{10}} = E(T|X = 1, z_1 = 0) = 20$
(4) $\frac{1}{\lambda_{11}} = E(T|X = 1, z_1 = 1) = 60.$

The survival time values are generated using the proportional hazards function $h(t, x, z) = h_0(t) \exp(\alpha_1 z_1 + \alpha_2 z_2 + \alpha_3 z_3 + \alpha_4 z_4 + \beta x)$, where $h_0 = 0.05$, $\alpha_1 = -\log(3)$, $\alpha_2 = 0$, $\alpha_3 = 0$, $\alpha_4 = 0$, and $\beta = 0$, whereas the survival time values are related only to $z_1$, which indicates by gender, and the other terms except for the $z_1$ term become zero. If the distribution corresponding to $h_0(t)$ conforms to the exponential distribution, then the hazard rate $h_0(t)$ is represented by $\lambda$.

As mentioned in Section 1, if more than one cause of death exists, cumulative incidence can be considered. In the case of two causes of death, we call the first one cause 1, and the other cause 2. Here, we consider the proportion of causes, as follows:

$$
P(C = 1|x) = \begin{cases} 0.6 & \text{if } x = 0 \\ 0.4 & \text{if } x = 1, \end{cases}
$$

where $C$ indicates the causes. Based on the scenario above, we perform simulation study. For each scenario, we generate two types of data: uncensored data, and censored data.

### 3.2 Comparison of performance between cumulative incidence and adjusted cumulative incidence with balanced data

In this section, we compare the performances of the regular cumulative incidence with those of the adjusted cumulative incidence under balanced data as well as unbalanced data. To calculate an adjusted Kaplan–Meier survival curve,

we first compute the weight $w_i$ by using a logistic regression. To obtain a weight $w_i$, we first calculate the conditional probability $\hat{p_i}$, which is the probability of being in the group given $z_i$, where i = 1, 2, 3, 4. Weight $w_i = \frac{1}{\hat{p_i}}$ is assigned to those subjects who belong to group 1, otherwise $w_i = \frac{1}{1-\hat{p_i}}$. By using equations (1), (2), (3), and (4), we compute the estimates of adjusted cumulative incidence curves. In addition, we calculate the regular cumulative incidence using the original version of equation (3).

We plot the estimates of adjusted Kaplan–Meier survival curves with the estimates of regular Kaplan–Meier survival curves under the balanced data set, the plots of which are shown in Appendix A in the Supplementary Material http://intlpress.com/site/pub/files/_supp/sii/2019/0012/0003/SII-2019-0012-0003-s002.pdf. Then, we plot the estimates of the regular cumulative incidence curves with the estimates of the adjusted cumulative incidence curves. Appendices B and C in the Supplementary Material show the regular cumulative incidence and the adjusted cumulative incidence for causes 1 and 2, respectively. For greater accuracy, we also show estimates of the regular cumulative incidence and the adjusted cumulative incidence for specific time for groups 0 and 1, which are shown in Table 1 of Appendix D in the Supplementary Material. Based on the result table, the differences between the unadjusted cumulative incidence and the adjusted cumulative incidence are negligible. That is, we can say that the unadjusted estimates and the adjusted estimates of cumulative incidence are very similar when data are balanced.

Next, we compare the estimates of the cumulative incidence curves with the estimates of the adjusted cumulative incidence curves in case that the data is censored, which indicates that the event has not happened during the analysis. For the censored data, we need to generate censoring times. The censored times, $C$, are generated from an exponential distribution with mean $\lambda_c$, where $\lambda_c$ follows a uniform$(0.02, 0.07)$ distribution. Then, we compare the created censoring times and the survival times that we generated in Section 3.1. If the censoring times are earlier than the survival time, then we assume the corresponding data is censored.

Based on the censored data, we plot the estimates of adjusted Kaplan–Meier survival curves (or the adjusted cumulative incidence curves) with the estimates of regular Kaplan–Meier survival curves (or the regular cumulative incidence curves), respectively. The related plots are shown in Appendix E, F, and G in the Supplementary Material. For greater accuracy, we also show estimates of the regular cumulative incidence and the adjusted cumulative incidence for specific times for groups 0 and 1, which are shown in Table 2 of Appendix D in the Supplementary Material.

To sum up, based on the results of Appendix D in the Supplementary Material, the cumulative incidence and the adjusted cumulative incidence show similar estimates under balanced data. It indicates that the adjusted cumulative incidence work equally well even though the regular cumulative incidence is designed under balanced data. In the next section, we compare those two approaches under unbalanced data.

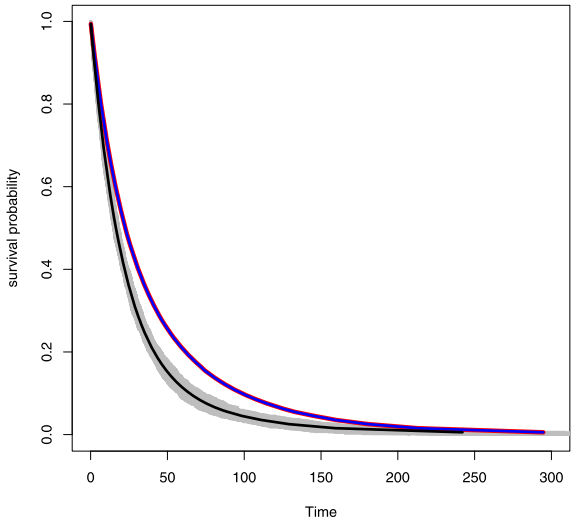
## 3.3 Comparison of performance between cumulative incidence and adjusted cumulative incidence for the unbalanced data

### 3.3.1 Uncensored data

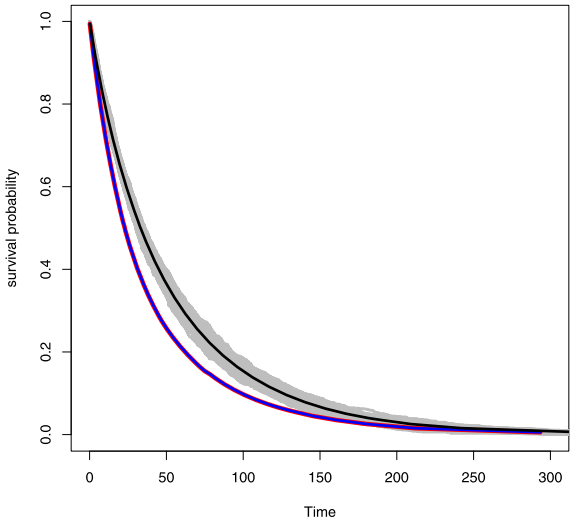
In this section, we compare the performance of the regular cumulative incidence with those of the adjusted cumulative incidence under unbalanced data. In Figure 1, we first plot the average of estimates of the adjusted Kaplan–Meier survival curves under the balanced data (shown as a red curve), the average of estimates of the regular Kaplan–Meier survival curves under the balanced data (shown as a blue curve) as refereces. Then, we plot the estimated Kaplan–Meier survival curves under the unbalanced data (shown as gray curves), and the average of estimates of the regular Kaplan–Meier curves under the unbalanced data (shown as a black curve). As shown in Figures 1 (a) and (b), the red curve and blue curve are similar. Nonetheless, the black curve in Figures 1 (a) and (b) is skewed far from the red and blue curve, which are the estimated survival function under balanced data. The black curve in Figure 1 (a) and (b) is skewed toward the survival function generated by exponential distributions with means of 60 and 20, respectively. This is because we assign 80% of the subjects to group 0, which has an average time of 60, and 20% of the subjects to group 1, which has an average time of 20. This is why the graphs shown in Figures 1 (a) and (b) are skewed.

Similarly, we compare the adjusted cumulative incidence with the regular cumulative incidence for Cause 1. In Figures 2, we plot the average of estimates of the adjusted cumulative incidence under the balanced data (shown as a red curve), and the average of estimates of the regular cumulative incidence under the true balanced data (shown as a blue curve) as references. Then, we shows the estimated cumulative incidence under the unbalanced data (shown as gray curves), and the average of estimates of the regular cumulative incidence under the unbalanced data (shown as a black curve). Figures 2 (a) and (b) show that the red curve and blue curve look similar. However, the black curve in Figures 2 (a) and (b) is skewed far from the cumulative incidence under balanced data, and skewed toward the cumulative incidence generated by exponential distributions with means of 60 and 20, respectively. Similar patterns are also observed for Cause 2 in Figure 6 (c) and (d), which show such skewness of estimates of the regular cumulative incidence under the unbalanced data.

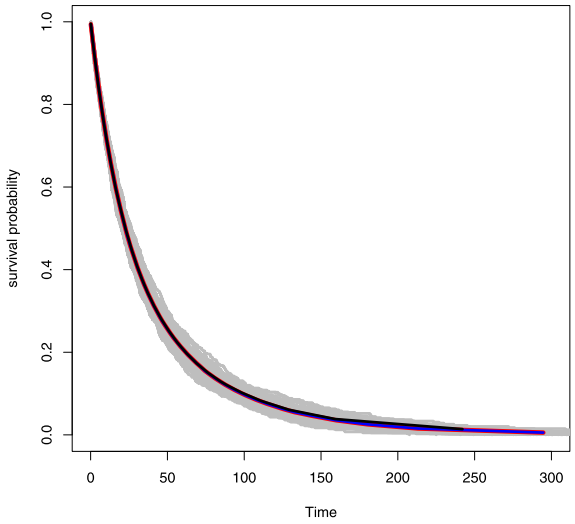
However, if we adjust the cumulative incidence curve against such a unbalanceness, the estimates are not skewed far from the true survival or cumulative incidence curves.



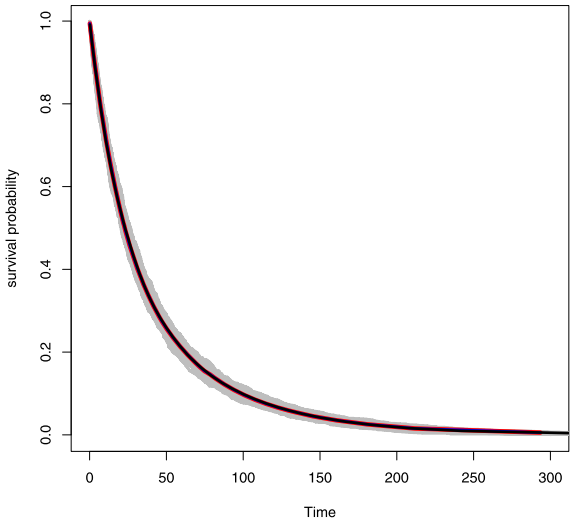

(a) Unadjusted Survival Function for Group X=0




(b) Unadjusted Survival Function for Group X=1



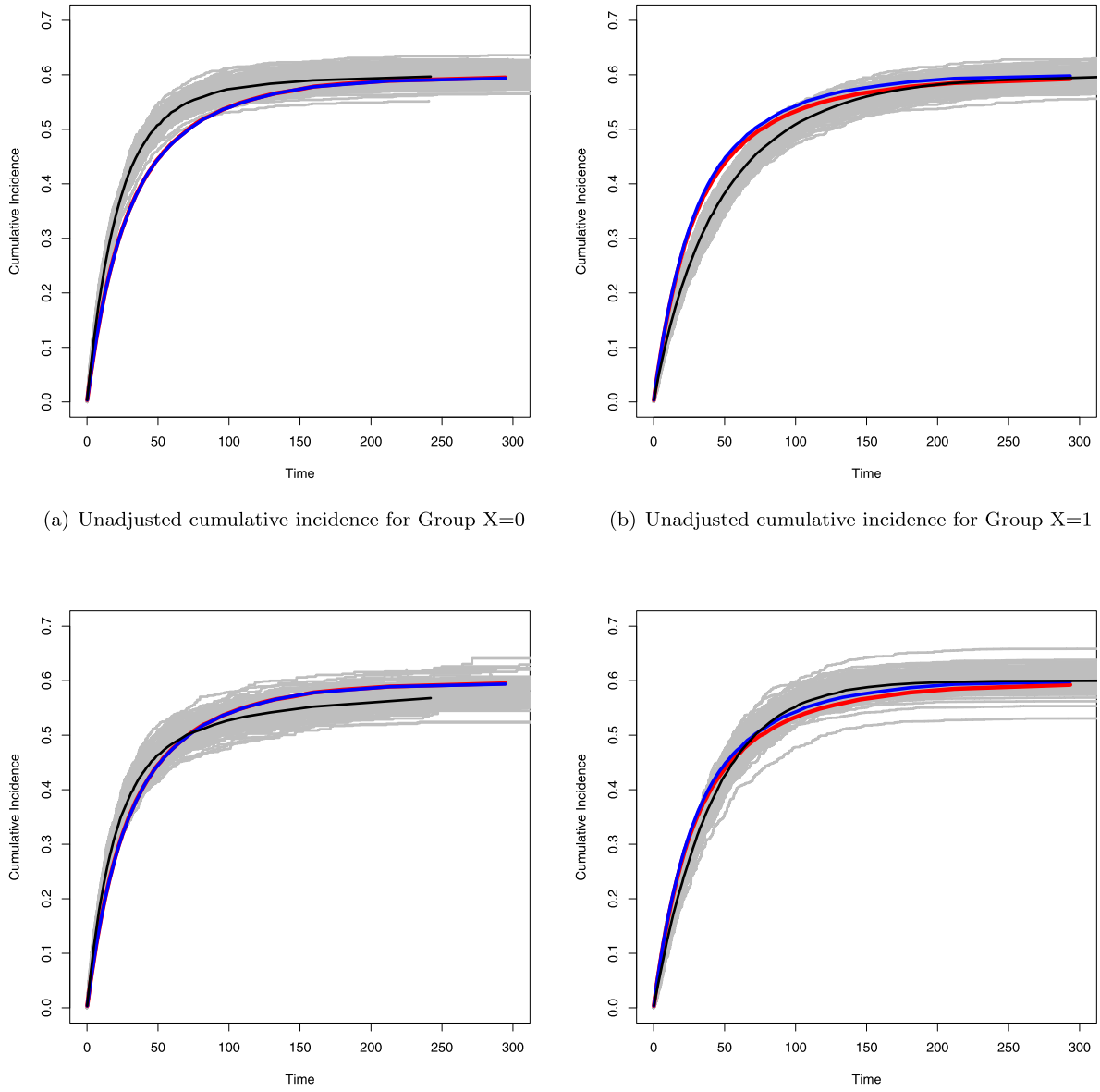

(c) Adjusted Survival Function for Group X=0




(d) Adjusted Survival Function for Group X=1

*Figure 1. Comparisons of regular Kaplan–Meier curves when unbalanced data are used. The average line of adjusted survival curves of balanced data (red curve); the average line of unadjusted survival curves of balanced data (blue curve); the average line of adjusted and unadjusted survival curves of unbalanced data (black curve); adjusted and unadjusted survival curves of unbalanced data (gray curves).*

Figures 1 (c) and (d) depict the adjusted Kaplan–Meier survival curves. We plot the average of estimates of adjusted survival curves under the balanced data (shown as red curves), and the average of estimates of the regular survival curves under the balanced data set (shown as a blue curve) as references. Then, we plot the average of estimates of the regular survival curves under the unbalanced data (shown as a black curve). The three curves look very similar. In addition, Figures 2 (c) and (d) present the adjusted cumulative incidence for Cause 1, and Figures 3 (c) and (d) show the adjusted cumulative incidence for Cause 2. In Figure 2 (c) and (d), the average of estimates of the adjusted cumulative incidence curve under the balanced data (shown as red curves), the average of estimates of the regular cumulative incidence curve under the balanced data (shown as a blue curve), and the average of estimates of the adjusted cumulative incidence curve of the unbalanced data (shown as a black curve) look very similar. The same pattern are




(a) Unadjusted cumulative incidence for Group X=0

(b) Unadjusted cumulative incidence for Group X=1

(c) Adjusted cumulative incidence for Group X=0

(d) Adjusted cumulative incidence for Group X=1

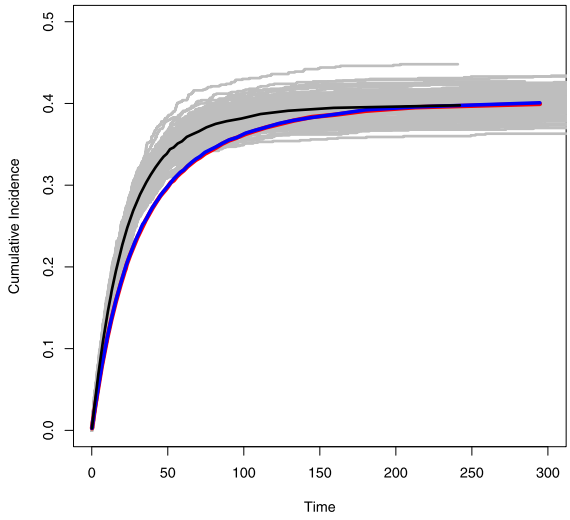
*Figure 2. Comparisons of regular cumulative incidence for unbalanced data and cause 1. The average line of adjusted cumulative incidence curves of balanced data (red curve); the average line of unadjusted cumulative incidence curves of balanced data (blue curve); the average line of adjusted and unadjusted cumulative incidence curves of unbalanced data (black curve); adjusted and unadjusted cumulative incidence curves of unbalanced data (gray curves).*

also shown in Figure 3 (c) and (d).

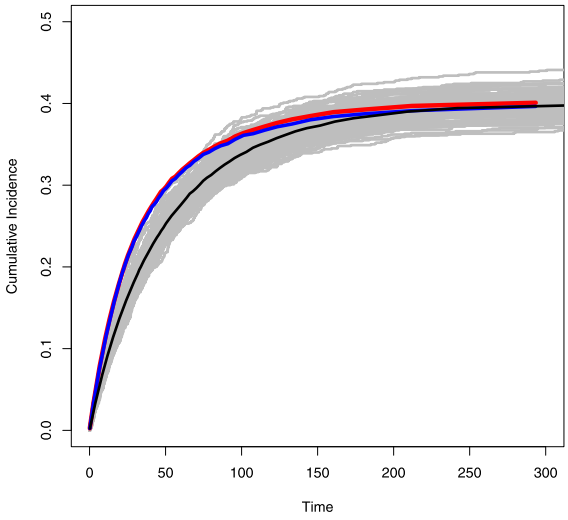
To sum up, if the data are unbalanced, the adjusted cumulative incidence approach provide estimated curves close to the curves estimated under balanced data. Thus, it reduces the effect of unbalanceness of data, and can capture true underlying survival incidence well. On the other hand, the regular cumulative incidence approach is relatively much sensitive to such effect of unbalanceness and provides biased estimates of survival incidence.

### 3.3.2 Censored data

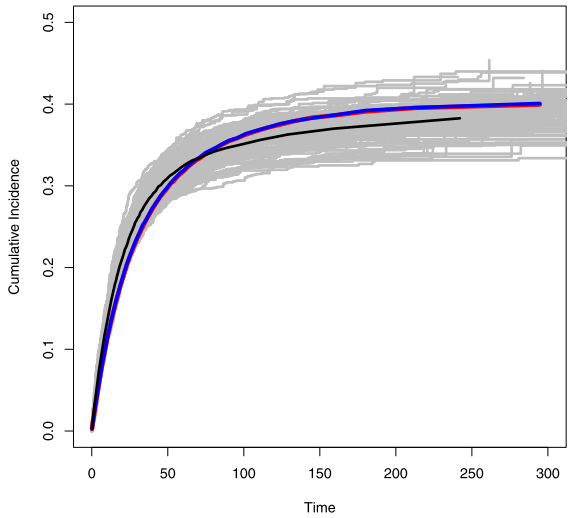
We also perform simulation studies under the unbalanced data with censoring. Figures 4 (a) and (b) depict estimates of the regular survival curves under unbalanced data, and Figures 4 (c) and (d) plot estimates of the adjusted survival curves under unbalanced data. Figures 5 (a) and (b) plot the unadjusted cumulative incidence for cause 1, and Figures 6 (a) and (b) show the estimates of regular cumulative incidence for cause 2. In addition, Figures 4 (c) and (d), 5 (c)



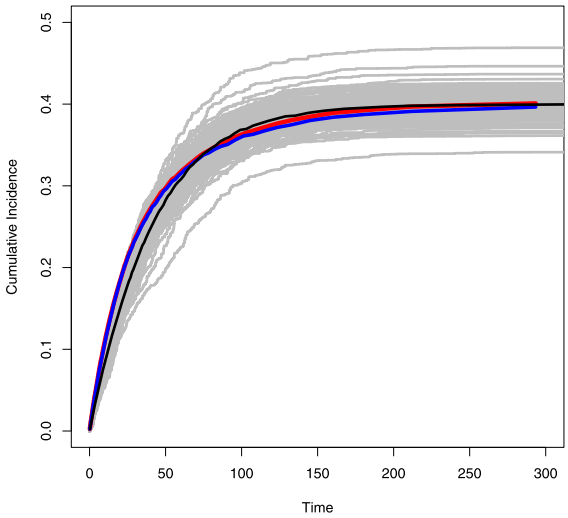

(a) Unadjusted cumulative incidence for Group X=0




(b) Unadjusted cumulative incidence for Group X=1




(c) Adjusted cumulative incidence for Group X=0




(d) Adjusted cumulative incidence for Group X=1

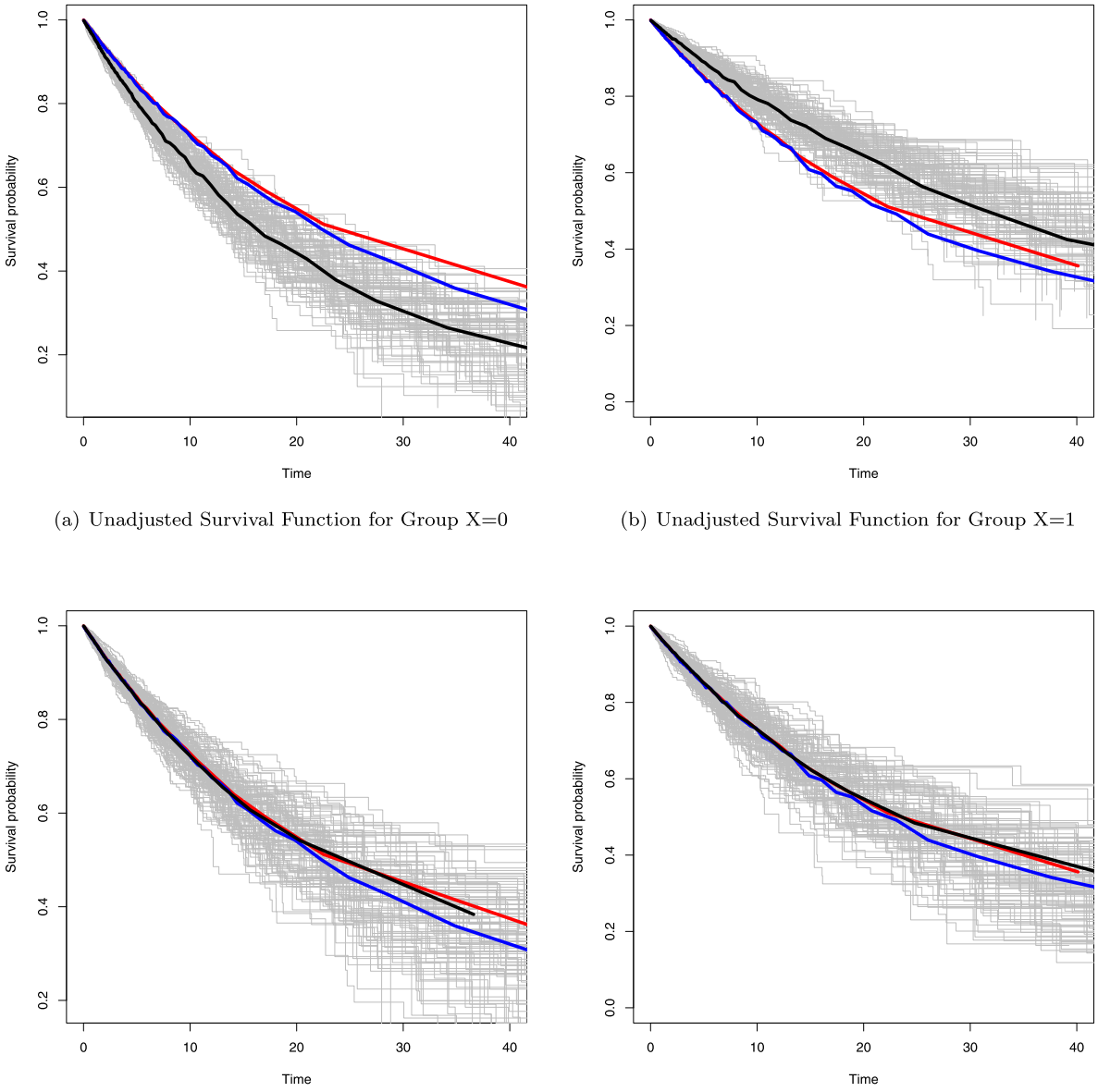
*Figure 3. Comparisons of adjusted cumulative incidence when unbalanced data are analyzed for cause 2. The average line of adjusted cumulative incidence curves of balanced data (red curve); the average line of unadjusted cumulative incidence curves of balanced data (blue curve); the average line of adjusted and unadjusted cumulative incidence curves of unbalanced data (black curve); the adjusted and unadjusted cumulative incidence curves of unbalanced data (gray curve).*

and (d), and 6 (c) and (d) show estimates of the adjusted survival curve, estimates of adjusted cumulative incidence for cause 1, and estimates of adjusted cumulative incidence for cause 2, respectively.

Similar to the result based on uncensored data, the estimates of the adjusted survival curves under unbalanced data are close to estimates of regular or adjusted survival curves under balanced data when the data is censored. However, the estimates of regular survival curves under unbalanced data are relatively far from the estimates of regular or adjusted survival curves under balanced data.

## 3.4 An adjusted Kaplan–Meier curve/cumulative incidence curve for data from a nonproportional hazards model

We also perform simulation study based on a nonproportional hazards model, where we generate the survival time




(a) Unadjusted Survival Function for Group X=0

(b) Unadjusted Survival Function for Group X=1

(c) Adjusted Survival Function for Group X=0

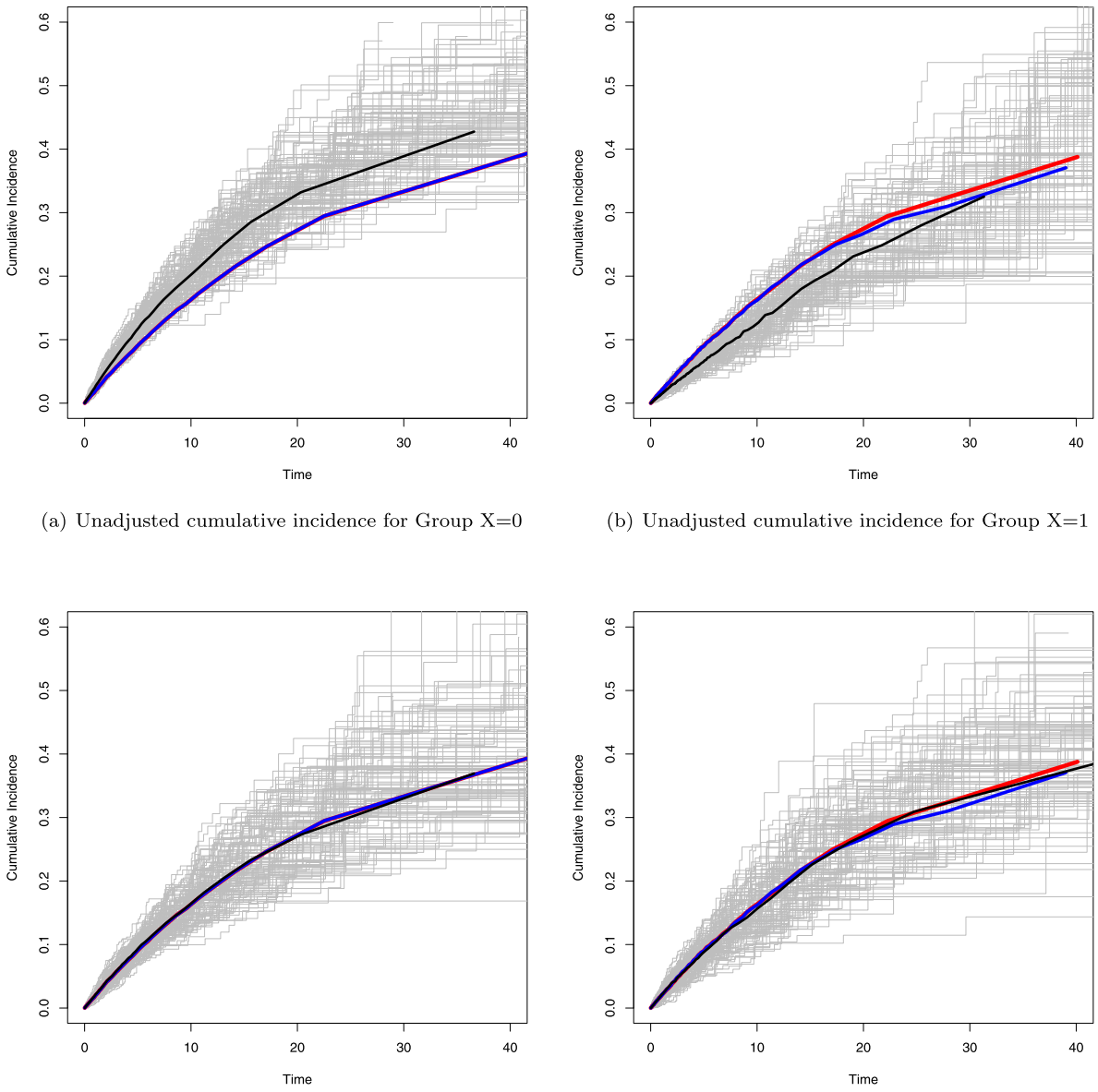
(d) Adjusted Survival Function for Group X=1

*Figure 4. Comparisons of Kaplan–Meier curves when unbalanced data are used. The average line of adjusted survival curves of balanced data (red curve); the average line of unadjusted survival curves of balanced data (blue curve); the average line of unadjusted survival curves of unbalanced data (black curve); the unadjusted survival curves of unbalanced data (gray curve).*

as follows: $\log(T) = \alpha_1 z_1 + \alpha_2 z_2 + \alpha_3 z_3 + \alpha_4 z_4 + \beta x + \varepsilon$, where $\alpha_1 = -\ln(3)$, $\alpha_2 = 0$, $\alpha_3 = 0$, $\alpha_4 = 0$ and $\beta = 0$. It is called an accelerated model, and an error term $\varepsilon$ follows a normal distribution with mean zero and standard deviation one. For the cumulative incidence, we consider the proportion of causes in the data set in Section 3.2, and other simulation setting are the same.

Under the data simulated from a nonproportional hazards model, we first compare the regular approach with the adjusted approach under unbalanced data with censoring. The plots for estimates of adjusted and regular survival curves are shown in Appendix H of the Supplementary Material. The plots for the cumulative incidence for causes 1 and 2 are presented in Appendices I and J of the Supplementary Material, respectively.

As shown in those figures above, the estimates of the adjusted cumulative incidence curves under unbalanced data are close to estimates of regular or adjusted cumulative incidence curves under balanced data. However, the estimates of regular cumulative incidence curves under unbal-




(a) Unadjusted cumulative incidence for Group X=0

(b) Unadjusted cumulative incidence for Group X=1

(c) Adjusted cumulative incidence for Group X=0

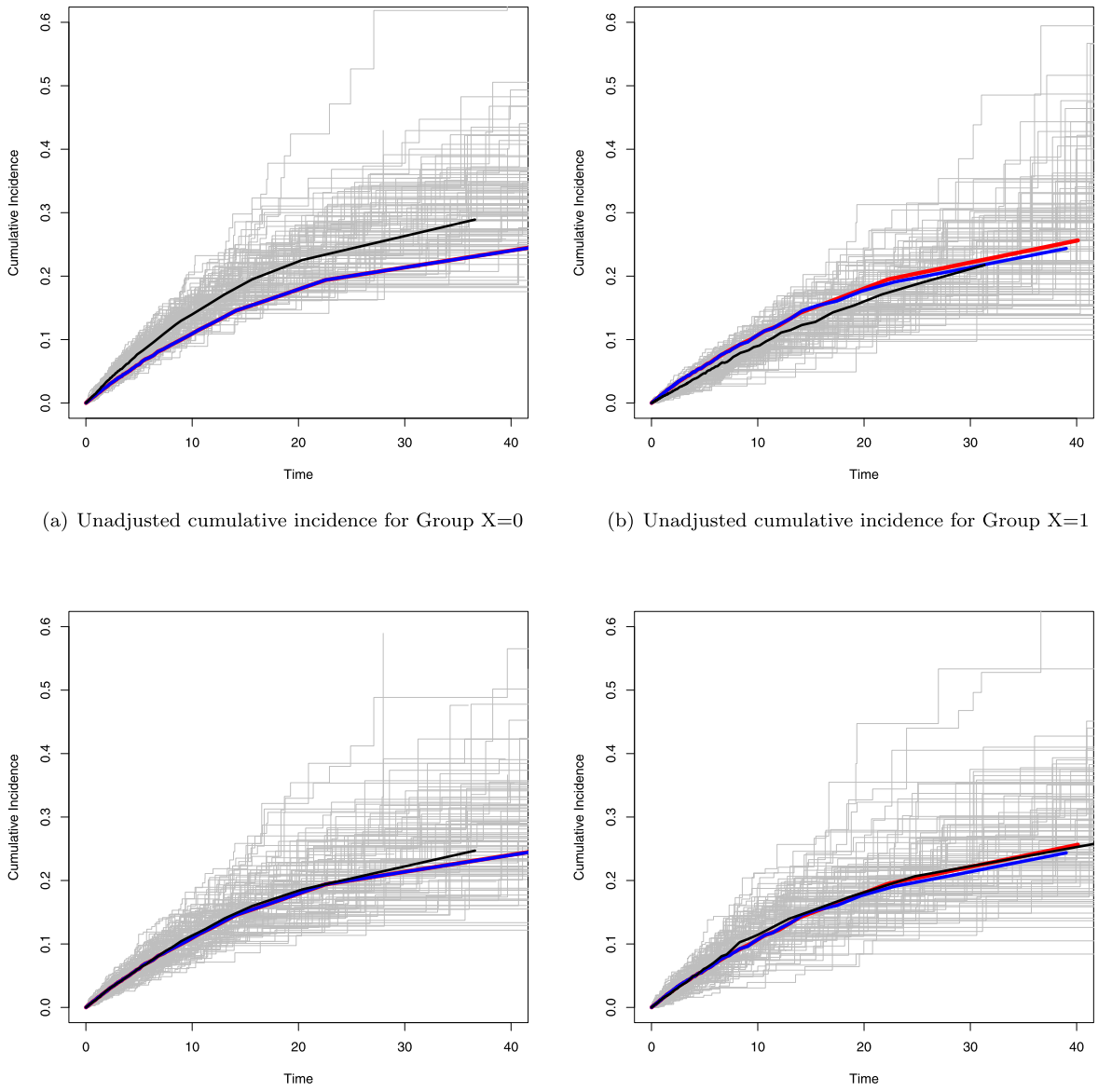
(d) Adjusted cumulative incidence for Group X=1

*Figure 5. Comparisons of cumulative incidence when unbalanced data are analyzed for cause 1. The average line of adjusted cumulative incidence curves of balanced data (red curve); the average line of unadjusted cumulative incidence curves of balanced data (blue curve); the average line of unadjusted cumulative incidence curves of unbalanced data (black curve); unadjusted cumulative incidence curves of unbalanced data (gray curves).*

anced data are relatively far from the estimates of regular or adjusted cumulative incidence curves under balanced data. Overall, the adjusted cumulative incidence approach is robust against data unbalanceness, but the regular cumulative incidence approach is sensitive.

## 4. PRACTICAL APPLICATION

In the following application, we use the cumulative incidence and adjusted cumulative incidence to compare the two estimators in a data set of 503 patients who have lower-grade glioma of the brain; this data set was obtained from TCGA (http://firebrowse.org). Glioma is a type of cancer that originates in the brain or spine. It has four stages, and stage number two is called brain lower-grade glioma. This stage occurs in young people 20–25 years of age and the affected cell types are astrocytes, oligodendroglioma, and oligoastrocytoma, which are called the "histological type" in these data. The treatments for this cancer




(a) Unadjusted cumulative incidence for Group X=0

(b) Unadjusted cumulative incidence for Group X=1

(c) Adjusted cumulative incidence for Group X=0

(d) Adjusted cumulative incidence for Group X=1

*Figure 6. Comparisons of cumulative incidence when unbalanced data are used for cause 2. The average line of adjusted cumulative incidence curves of balanced data (red curve); the average line of unadjusted cumulative incidence curves of balanced data (blue curve); the average line of unadjusted cumulative incidence curves of unbalanced data (black curve); unadjusted cumulative incidence curves of unbalanced data (gray curves).*

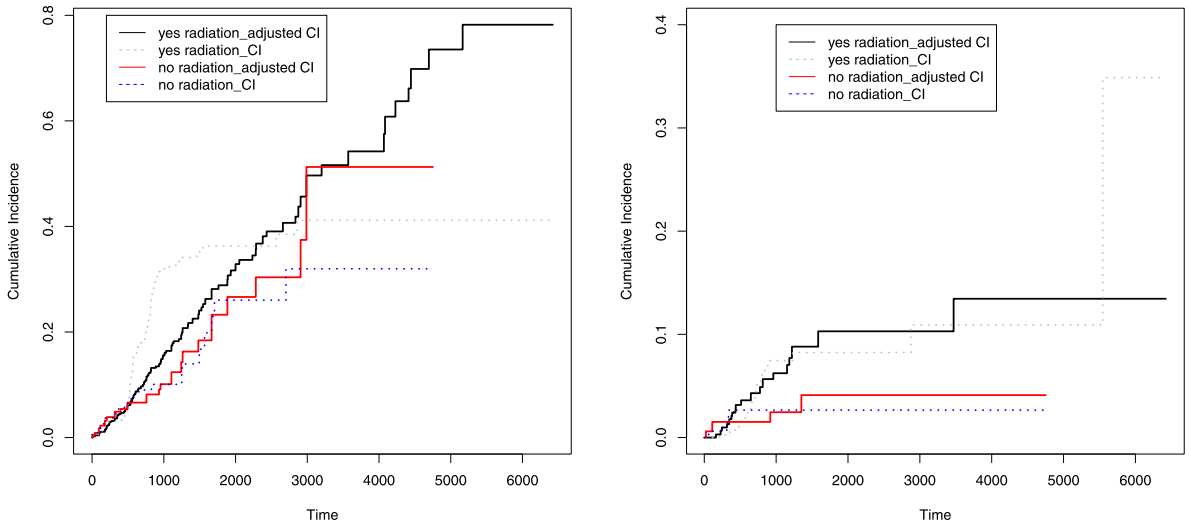
are radiation therapy or a surgical treatment, depending on the tumor characteristics (American Brain Tumor Association, http://www.abta.org/brain-tumor-information/types-of-tumors.glioma.html). The observed variables in the data set are the patient's status: whether he/she has died, whether he/she received radiation therapy, whether he/she was decimated by the tumor, and certain gene information related to the brain lower-grade glioma.

The purpose of this practical application is to confirm that the adjusted cumulative incidence is different to the regular cumulative incidence. In this example, we divide the group based on the variable "radiation therapy," which indicates whether or not patients underwent radiation therapy. In this application, the treatment is "radiation therapy."

The variable "histological type" indicates whether the patients received radiation therapy. In the group of patients, the proportion of the histological type in each group based on radiation therapy is presented in Table 1. As mentioned earlier, to adjust each estimator, we have to set the weights. The response variable "radiation therapy" is regressed on

*Table 1. The proportion of a histological type in each group based on radiation therapy*

| Histological type | astrocytoma | oligoastrocytoma | oligodendroglioma |
|---|---|---|---|
| Radiation therapy | 45.3% | 24.9% | 29.8% |
| No radiation therapy | 24.0% | 28.4% | 47.5% |




(a) Adjusted cumulative incidence and cumulative incidence in a real study (cause 1)

(b) Adjusted cumulative incidence and cumulative incidence in a real study (cause 2)

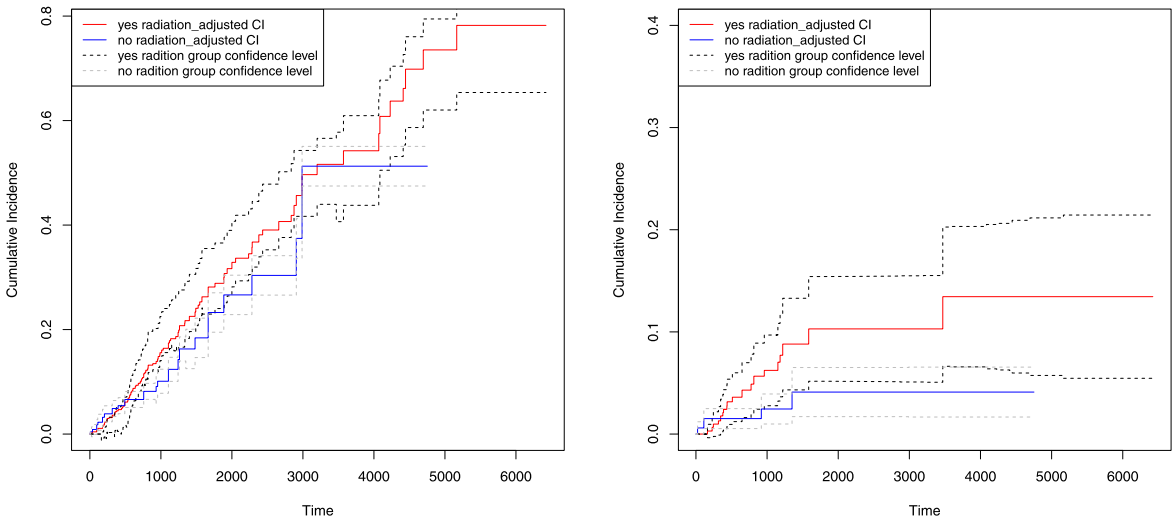
*Figure 7. Comparison of adjusted cumulative incidence and cumulative incidence in a real study. Cumulative incidence for the "yes" radiation group (gray dotted curve); adjusted cumulative incidence for the radiation group (black curve); cumulative incidence for the "no" radiation group (blue dotted curve); the adjusted cumulative incidence for the "no" radiation group (red curve).*

several variables that have a significant effect on the "radiation therapy" variable. The variables analyzed in order to calculate the weights are histological type and several important genes that are related to brain lower-grade glioma. After we estimate the weights, we obtain the regular cumulative incidence and the adjusted cumulative incidence. The resulting plots for causes 1 and 2 are presented in Figure 7. As shown in the figure, the regular cumulative incidence data for the group who received radiation therapy (gray dotted curve) and the group who did not receive radiation therapy (blue dotted curve) do not lie close together. Nevertheless, the adjusted cumulative incidence for the group who received radiation therapy (black curve) and the adjusted cumulative incidence for the group who did not receive radiation therapy (red curve) are close. This means that even if we have confounding variables, such as "histological type" in this study, we obtain similar estimated curves between multiple unbalanced treatments after weighting. We calculate the pointwise confidence level using Equation (8) and plot the confidence level for the adjusted estimators for causes 1 and 2. The plots for causes 1 and 2 are shown in Figures 8 (a) and (b), respectively.

## 5. CONCLUSIONS

In this study, we explained two estimators, regular estimator and adjusted estimator, based on the proportional hazards model using unbalanced data and balanced data. Using the simulation proposed by Xie and Liu (2005), we compared the performance of the adjustments of the two estimators. For the balanced data, we showed that with the adjustment and without the adjustment, they converge in the same way. We found that the performance of the adjusted estimators when analyzing unbalanced data is robust against unbalanceness. A regular cumulative incidence is sensitive to such unbalanceness, so an adjusted cumulative incidence often shows better performance. Overall, the adjusted cumulative incidence is preferred.

Other extensions of the adjusted cumulative incidence are possible, for example, if we want to verify the adjusted cumulative incidence for high-dimensional data. Particularly in biological data, high-dimensional data have far more variables than the sample size. This approach requires additional work in biometrics. In addition, we want to apply this estimator to another type of nonparametric




(a) Cumulative incidence for both groups and its 95% confidence interval (cause 1)

(b) Cumulative incidence for both groups and its 95% confidence interval (cause 2)

*Figure 8. Cumulative incidence for both groups and its 95% confidence interval. Estimated cumulative incidence (bold curves); 95% pointwise confidence interval (dashed curves).*

method, namely the Cox proportional hazards regression model.

## APPENDIX A. ASYMPTOTIC VARIANCE ESTIMATION OF ACIF

Using arguments similar to those of Choudhury (2002) and Xie and Liu (2005), it can be shown that for $t_j \le t < t_{j+1}$, the estimated variance of $\hat{F}_{11}^w$ (i.e., $\widehat{\text{var}}[\hat{F}_{11}^w(t)]$) is equal to

$$(9)\qquad \sum_{l}^{j} \widehat{\text{var}}[\hat{\lambda}_{1,l1}^w \hat{S}_1^w(t_{l-1})] + 2\sum_{l}^{j-1}\sum_{s=l+1}^{j} \widehat{\text{cov}}[\hat{\lambda}_{1,l1}^w \hat{S}_1^w(t_{l-1}), \hat{\lambda}_{1,s1}^w \hat{S}_1^w(t_{s-1})],$$

where

$$\widehat{\text{var}}[\hat{\lambda}_{1,l1}^w \hat{S}_1^w(t_{l-1})] = \{\hat{\lambda}_{1,l1}^w \hat{S}_1^w(t_{l-1})\}^2 \left\{ \frac{1-\hat{\lambda}_{1,l1}^w}{M_l \hat{\lambda}_{1,l1}^w} + \sum_{m=1}^{l-1} \frac{1-\hat{\theta}_{m1}^w}{M_m \hat{\theta}_{m1}^w} \right\},$$

and

$$\widehat{\text{cov}}[\hat{\lambda}_{1,l1}^w \hat{S}_1^w(t_{l-1}), \hat{\lambda}_{1,s1}^w \hat{S}_1^w(t_{s-1})] = \{\hat{\lambda}_{1,l1}^w \hat{S}_1^w(t_{l-1}) \hat{\lambda}_{1,s1}^w \hat{S}_1^w(t_{s-1})\} \times \left\{ \left(1 - \frac{1}{Y_{l1}^w}\right) \prod_{m=1}^{l-1} \left(\frac{1-\hat{\theta}_{m1}^w}{M_m \hat{\theta}_{m1}^w}\right) - 1 \right\}.$$

Using a Taylor series linear approximation, we have

$$\prod_{m=1}^{l-1} \left(\frac{1-\hat{\theta}_{m1}^w}{M_m \hat{\theta}_{m1}^w}\right) \approx 1 + \sum_{m=1}^{l-1} \left(\frac{1-\hat{\theta}_{m1}^w}{M_m \hat{\theta}_{m1}^w}\right),$$

and, under condition (5), we have $M_m^{-1} \to 0$ and, therefore, if we ignore the terms of a smaller order of magnitude than $M_m^{-2}$, then equation (9) is reduced to

$$\begin{aligned}\widehat{\text{var}}[\hat{F}_{11}^w(t)] &= \sum_{l=1}^{j} \{\hat{\lambda}_{1,l1}^w \hat{S}_1^w(t_{l-1})\}^2 \left\{ \frac{1-\hat{\lambda}_{1,l1}^w}{M_l \hat{\lambda}_{1,l1}^w} + \sum_{m=1}^{l-1} \frac{1-\hat{\theta}_{m1}^w}{M_m \hat{\theta}_{m1}^w} \right\} \\ &\quad + 2\sum_{l=1}^{j-1}\sum_{s=l+1}^{j} \{\hat{\lambda}_{1,l1}^w \hat{S}_1^w(t_{l-1}) \hat{\lambda}_{1,s1}^w \hat{S}_1^w(t_{s-1})\} \\ &\quad \times \left\{ -\frac{1}{Y_{l1}^w} + \sum_{m=1}^{l-1} \frac{1-\hat{\theta}_{m1}^w}{M_m \hat{\theta}_{m1}^w} \right\}.\end{aligned}$$

## ACKNOWLEDGEMENTS

This research was supported by a grant from the National Research Foundation of Korea (NRF-2017R1E1A1A03070507, NRF-2017R1C1B1004817) and by a Korea University Grant (K1719881, K1607341, K1822881). Please send any inquiries to the corresponding authors, Sung Won Han (School of Industrial Management Engineering, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul, 02841, Republic of Korea. E-mail: swhan@korea.ac.kr), and Eun-Seok Ryu (Depart-

ment of Computer Engineering, Gachon University, 1342, Seongnam-daero, Sujeong-gu, Seongnam-si, Gyeonggi-do, 13120, Republic of Korea. E-mail: esryu@gachon.ac.kr).

*Received 13 April 2017*

## REFERENCES


Amato, D. A. (1988). A generalized Kaplan-Meier estimator for heterogeneous populations. *Commun. Stat. Theory Methods* **17** 263–286. MR0940978

Andersen, P. K., Geskus, R. B., De Witte, T., and Putter, H. (2012). Competing risks in epidemiology: Possibilities and pitfalls. *Int. J. Epidemiol.* **41**(3) 861–870.

Antolini, L., Biganzoli, E. M., and Boracchi, P. (2006). Crude cumulative incidence in the form of a Horvitz-Thompson like and Kaplan-Meier like estimator. COBRA Preprint Series. 10, http://biostats.bepress.com/cobra/art10.

Benichou, J. and Gail, M. H. (1990). Estimates of absolute cause-specific risk in cohort studies. *Biometrics* **46**(3) 813–826.

Choi, S. and Huang, X. (2014). Maximum likelihood estimation of semiparametric mixture component models for competing risks data. *Biometrics* **70** 588–598. MR3261778

Choudhury, J. B. (2002). Non-parametric confidence interval estimation for competing risks analysis: Application to contraceptive data. *Stat. Med.* **21** 1129–1144.

Cupples, L. A., Gangnon, D. R., Ramaswamy, R., and D'Agostino, R. B. (1995). Age-adjusted survival curves with application in the Framingham study. *Stat. Med.* **14** 1731–1744.

Dignam, J. J. and Kocherginsky, M. N. (2008). Choice of interpretation of statistical tests used when competing risks are present. *J. Clin. Oncol.* **26**(24) 4027–4034.

Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *J. Am. Stat. Assoc.* **94**(446) 496–509. MR1702320

Freidlin, B. and Korn, E. L. (2005). Testing treatment effects in the presence of competing risks. *Stat. Med.* **24** 1703–1712. MR2137645

Gray, R. J. (1988). A class of K-sample tests for comparing the cumulative incidence of a competing risk. *Ann. Statist.* **16**(3) 1141–1154. MR0959192

Gail, M. H. and Pfeiffer, R. M. (2005). On criteria for evaluating models of absolute risk. *Biostatistics* **6**(2) 227–239.

Hankey, B. F. and Myers, M. H. (1971). Evaluating differences in survival between two groups of patients. *J. Chron. Dis.* **24** 523–531.

Kalbfleisch, J. D. and Prentice, R. L. (2002). The Statistical Analysis of Failure Time Data. 2nd Ed. John Wiley, New York, NY. MR1924807

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* **53** 457–481. MR0093867

Kim, H. T. (2007). Cumulative incidence in competing risks data and competing risks regression analysis. *Clin. Cancer. Res.* **13** 559–565.

Klein, J. P. and Moeschberger, M. L. (2003). Survival Analysis: Techniques for Censored and Truncated Data. Springer-Verlag, New York.

Latouche, A. and Porcher, R. (2007). Sample size calculations in the presence of competing risks. *Stat. Med.* **26** 5370–5380. MR2416833

Lau, B., Cole, S. R., and Gange, S. J. (2009). Competing risk regression models for epidemiologic data. *Am. J. Epidemiol.* **170** 244–256.

Lin, D. Y. (1997). Nonparametric inference for cumulative incidence functions in competing risks studies. *Stat. Med.* **16** 901–910.

Nieto, F. J. and Coresh, J. (1996). Adjusting survival curves for confounders: A review and a new method. *Am. J. Epidemiol.* **143** 1068–1069.

Peng, L. and Fine, J. P. (2009). Competing risks quantile regression. *J. Am. Stat. Assoc.* **104** 1440–1453. MR2596999

Pepe, M. S. and Mori, M. (1993). Kaplan-Meier, marginal, or conditional probability curves in summarizing competing risks failure time data. *Stat. Med.* **12** 737–751.

Pintilie, M. (2002). Dealing with competing risks: Testing covariates and calculating sample size. *Stat. Med.* **21** 3317–3324.

Prentice, R. L., Kalbfleisch, J. D., Peterson, A. V., Flournoy, N., Farewell, V. T., and Breslow, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics* **34** 541–554. MR0570114

Rebora, P., Antolini, L., Glidden, D. V., and Valsecchi, M. G. (2016). Crude incidence in two-phase designs in the presence of competing risks. *BMC Med. Res. Methodol.* **16** 5.

Robins, J. M. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In A IDS Epidemiology-Methodological Issues, Jewell, N., Dietz, K., and Farewell, V. (eds.), 297–331. Boston: Birkhauser.

Satagopan, J. M., Ben-Porat, L., Berwick, M., Robson, M., Kutler, D., and Auerbach, A. D. (2004). A note on competing risks in survival data analysis. *Br. J. Cancer* **91**(7) 1229–1235.

Wang, C. Y., Wang, S., Zhao, L. P., and Ou, S. T. (1997). Weighted semiparametric estimation in regression analysis with missing covariate data. *J. Amer. Statist. Assoc.* **92** 512–525. MR1467845

Xie, J., and Liu, C. (2005). Adjusted Kaplan Meier estimator and log-rank test with inverse probability of treatment weighting for survival data. *Statist. Med.* **24**(20) 3089–3110.



Sangbum Choi
Department of Statistics
Korea University
145 Anam-ro, Seongbuk-gu
Seoul 136-713
Republic of Korea
E-mail address: choisang@korea.ac.kr

Chaewon Kim
School of Industrial Management Engineering
Korea University
145 Anam-Ro, Seongbuk-Gu
Seoul 02841
Republic of Korea
E-mail address: ckim92@korea.ac.kr

Hua Zhong
Division of Biostatistics
Department of Population Health
New York University School of Medicine
650 First Avenue, Room 540
New York, NY 10016
USA
E-mail address: judy.zhong@nyumc.org

Eun-Seok Ryu
Department of Computer Engineering
Gachon University
1342, Seongnam-daero, Sujeong-gu, Seongnam-si
Gyeonggi-do, 13120
Republic of Korea
E-mail address: esryu@gachon.ac.kr

Sung Won Han
School of Industrial Management Engineering
Korea University
145 Anam-Ro, Seongbuk-Gu
Seoul 02841
Republic of Korea
E-mail address: swhan@korea.ac.kr