LETTERS IN BIOMATHEMATICS
An International Journal

RESEARCH ARTICLE

OPEN ACCESS

Comparison of Regression Approaches for Analyzing Survival Data in the Presence of Competing Risks: An Application to COVID-19

Sarada Ghosh[a], G. P. Samanta[a], Anuj Mubayi[b,c]

[a]Department of Mathematics, Indian Institute of Engineering Science and Technology, Shibpur, Howrah-711103, India; [b]College of Health Solutions, Arizona State University, Tempe, USA; [c]Department of Mathematics, Illinois State University, Normal, USA

ABSTRACT

Emerging infectious diseases have impacted human race regularly with the past few decades alone has been rife with outbreaks such as H7N9 Avian-influenza (2013), Ebola (2014), MERS-CoV (2012), SARS-CoV1 (2003), and Zika (2015). COVID-19 coronavirus variants are emerging across the globe causing ongoing pandemic. Older age, male sex, number of comorbidities, and access to timely health care are identified as some of the risk factors associated with COVID-19 mortality. The regression approaches for capturing the competing risks are applied to COVID-19 in this work. The most commonly used approaches are the cause-specific and sub-distribution hazards regression which are applied on the COVID-19 incidence-data from USA. Additionally, the pseudo-observation approach, which allows for analysis of survival data, is applied on the same data. The simulations are carried out to compare approaches under different scenarios and also illustrate the relative effect of COVID-19 infected people based on their gender and age.

ARTICLE HISTORY
Received May 7, 2020
Accepted January 27, 2021

KEYWORDS
Competing risk, Cause-specific hazard regression, Sub-distribution hazard regression, Cumulative incidence function, Pseudo approach

1 Introduction

Severe Acute Respiratory Syndrome (SARS), a viral respiratory disease, is the first major novel emerging infectious disease that originated in southern China in November 2002 and hit the international community in the 21st century. It reached Hong Kong in February 2003 and spread rapidly thereafter to 29 countries/regions on five continents, infecting the global cumulative total of 8098 infected persons with 774 deaths during the outbreak (Lam et al., 2003). No cases of SARS (now referred as CoV-1) have been reported worldwide since 2004. The related virus SARS coronavirus (named as SARS-CoV-2) is the cause of the ongoing 2019 coronavirus pandemic. On 31 December 2019, the outbreak has been traced to a novel strain of coronavirus (known as COVID-19), giving the interim name 2019-nCoV by WHO but later it is renamed as SARS-CoV-2 by the International Committee on Taxonomy of Viruses. Some researchers have suggested that the Huanan Seafood Wholesale Market, Wuhan may not be the original source of viral transmission to humans (Park, 2020). The virus primarily spreads among people via exhaled respiratory droplets such as coughing or sneezing. The World Health Organization has declared the situation a pandemic with some serious travel restrictions. Epidemiology of SARS-CoV-2 has been found to be different than of SARS-CoV-1. In COVID-19 patients, the time between exposure and symptom onset is estimated to be around five days, but may range from two to fourteen days. Among those who died from the disease, the time from development of symptoms to death is between 6 to 41 days, with a median of 14 days. Most of the people who died were elderly: (i) about 80% of deaths were in those over 60 and (ii) 75% had pre-existing health conditions including cardiovascular diseases and diabetes. By the end of 2019, a novel coronavirus that was originated from Wuhan, a city in China, has caused more than 90 million cases and 2 million deaths worldwide. Coupled with the economic costs resulting from restriction of movement of individuals, the pandemic has highlighted the necessity for a rapid coordinated international response to disease control. Coronaviruses vary significantly in risk factors and first step is to identify cause of high incidence and mortality in some of the regions.

The analysis of data from COVID-19 patients is essential to understand the clinical prognosis, to develop potential therapeutic agents for novel pathogen and to design intervention strategies such as for vaccine implementation. However, many

CONTACT Anuj Mubayi anujmubayi@yahoo.com

studies on COVID-19 have investigated mortality and incidence data using survival models without considering the presence of competing risks. Hence, for given clinical data, there is a need to identify appropriate statistical models which are necessary to analyze complex clinical information. Often either competing events are ignored or inappropriate regression based statistical methods are used in time-to-event analysis. However, these events can be extremely informative in understanding impact of risk factors under certain circumstances. Standard logistic regression is a popular tool for examining associations between risk factors and the event of interest if patient data are available.

In this study, the competing risks framework is described and problems occurring from analysis of event time data are presented. The goal of the research is to compare and contrast selected regression approaches for competing risks using COVID-19 reported data from USA. In particular, cause-specific hazards, sub-distribution hazards, and the pseudo-observation approaches are used to perform comparative analysis of survival data. The competing risks setting is described in Section 3, including discussions on different views on the competing risks situation and on the non-identifiability problem, and presentation of relevant terms/quantities used for description of competing risks data. In Section 4, regression models for the competing risks setting, which are proposed in the literature, are described and compared regarding model assumptions, applicability, and interpretation of obtained results. Various extensions of the basic models are mentioned and literature for further reading is given. A special focus lies on the derivation of estimates for cause-specific and sub-distribution hazard rates and also deduce the regression models based on pseudo-observations. In Section 4, various simulation studies are performed to estimate the cause-specific and sub-distribution hazard rates and hazard ratios. In Section 5, applications of the presented methods for competing risks regression are illustrated with real life examples on COVID-19. Finally, the presented methods as well as the findings from the simulation study and the data analyses are discussed in the last section.

2 Methods and Materials

2.1 Data Sources

The data of novel coronavirus (COVID-19) are obtained from Kaggle.com website (Kaggle, 2019, accessed in March–April 2020). The website contains the number of new cases that are being reported daily from different countries (that is, Italy, Spain, Germany, USA etc.) around the world. The dataset also has information about 50 U.S. states and as a case study we have considered data from the first two months of reported COVID-19 cases until the third week of April 2020. In this work, at first we compare the mortality rate among USA and Italy choosing as two groups and next we fit the underlying models to the data set of prevalence due to COVID-19 in the New York city where the highest number of cases were reported from U.S.

2.2 Competing Risks Preliminaries

In survival analysis, competing risks occur frequently. A competing risk is an event whose occurrence precludes the occurrence of the primary event of interest. In a study examining time to death attributable to cardiovascular causes, death attributable to non-cardiovascular causes is a competing risk (Austin et al., 2016). When an individual is under risk of failing from K distinct types of event, these different event types are called competing risks which are broadly covered in the statistical literature (Beyersmann et al., 2012). An alternative approach to competing risks is consideration of a bivariate random variable (T, D), where T is a random variable for the event time and D is a random variable for the event type. The competing risks process can then be interpreted as a special case of a multi-state model (Andersen and Keiding, 2012), leading to the intuitive definitions of cumulative incidence functions and cause-specific hazard rates. For each individual, $i = 1, 2, \dots, n$, the couple of event time or last time known to be free of any event t_i and a status variable indicating the type of event $d_i \in \{1, 2, \dots, K\}$ or a censored event time $(d_i = 0)$ is observed. As in the competing risks setting individuals can fail from different event types, measures used for standard survival analysis with only one certain type of event have to be adapted. In this section the most important and commonly used concepts and quantities have been presented.

2.2.1 Cumulative Incidence Function (CIF) of Occurrence of k^{th} Type

The probability for occurrence of every event type k out of the possible event types $1, \dots, K$ up to a given time t can be described in the presence of competing risks. That probability is mostly known as "cumulative incidence function" (CIF) for event type k in the literature. In this work, symbolically it is denoted and defined as

$$\bar{F}_k(t) = P(T \leq t, D = k) \tag{1}$$

where T is a strictly positive random variable corresponds to the "event time" and D is a random variable for the "type of event". Apart from this, there exists some another important names in statistical field known as "crude event probability" (Tsiatis, 2005; Lambert et al., 2010) or "sub-distribution function" (Resche-Rigon and Chevret, 2006; Pintilie, 2007). The name "sub-distribution function" is motivated by the fact that $\bar{F}_k(t)$ is not a real distribution function since it does not converge to one as t tends to infinity. But to the overall probability for an event of type k,

$$\lim_{t \to \infty} \bar{F}_k(t) = P(D = k). \tag{2}$$

For a given time t, the CIF of all K event types sum up to one minus the probability of being event-free up to time t, which is sometimes called the "overall survivor function". Symbolically it is denoted and defined as

$$S_{osf}(t) = 1 - \sum_{k=1}^{K} \bar{F}_k(t)). \tag{3}$$

$$\therefore \lim_{t \to \infty} \sum_{k=1}^{K} \bar{F}_k(t) = 1. \tag{4}$$

To estimate the CIF for event type k, the "cause-specific hazard function" is introduced in the next section.

2.2.2 Cause-specific Hazard Rate of k^{th} Type

In the field of survival analysis, hazard rates play an uttermost important role for analysis of competing risks dataset. The "cause-specific hazard rate" for event type k is the natural adaptation of the common hazard rate providing an individual's probability for failing from an event of type k in an infinitesimal small time interval t to $t + \Delta t$ given the individual did not fail from any event up to time t. Symbolically it is denoted and defined as

$$\lambda_k(t) \equiv \lambda_k^{cs}(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t, D = k \mid T \geq t)}{\Delta t}. \tag{5}$$

Considering mutually exclusive terminal events, the cause-specific hazards for all K event types at time t sum up to the overall hazard rate for failing from any event at t:

$$\lambda_{osf}(t) = \sum_{k=1}^{K} \lambda_k(t). \tag{6}$$

In analogy to standard survival analysis the cumulative cause-specific hazard rate for event type k at time t is the integral over the cause-specific hazard function from time 0 to t:

$$\Lambda_k(t) = \int_0^t \lambda_k(s)\, ds. \tag{7}$$

The "overall survivor function" $S_{osf}(t)$ denoting the probability of being free from any event up to time t, depends on the (cumulative) cause-specific hazard functions for all K types of event, which sum up to the overall (cumulative) hazard rate:

$$S_{osf}(t) = \exp\left(- \sum_{k=1}^{K} \lambda_k(t) \right) = \exp(-\Lambda_{osf}(t)). \tag{8}$$

The relationship between the CIF for event type k and the cause-specific hazard functions can be expressed as

$$\bar{F}_k(t) = \int_0^t \lambda_k(s) S_{osf}(s) ds = \int_0^t \lambda_k(s) \exp\left(- \sum_{l=1}^{K} \lambda_l(s) \right) ds. \tag{9}$$

As can be seen from (9), the CIF for event type k depends on the cause-specific hazard functions for all K types of event, indicating that risks for all event types have an effect on the probability for an event of type k.

2.2.3 Sub-distribution Hazard Rate of k^{th} Type

In the presence of competing risks, Gray (1988) introduced the sub-distribution hazard rate for event type k, denoted as $\gamma_k(t)$, which differs from the cause-specific hazard rate shown in (5) by the definition of its risk set. For the sub-distribution hazard rate for event type k at time t, individuals that failed from an event other than k prior to t remain in the risk set. The underlying hazard rate is defined as follows:

$$\gamma_k(t) \equiv \gamma_k^{sd}(t) = \lim_{\Delta t \to 0} \frac{P\big(t \leq T < t + \Delta t, D = k \mid \{T \geq t\} \cup \{T < t, D \neq k\}\big)}{\Delta t}. \tag{10}$$

The link between the CIF and the sub-distribution hazard is as follows:

$$\bar{F}_k(t) = 1 - \exp\left(-\Gamma_k(t)\right) \tag{11}$$

with $\Gamma_k(t)$ denoting the cumulative sub-distribution hazard

$$\Gamma_k(t) = \int_0^t \gamma_k(s)\, ds. \tag{12}$$

Competing events do not have to be accounted explicitly since these are considered implicitly in the adapted risk set. As $\gamma_k(t)$ provides the properties of a hazard rate for the sub-distribution function $\bar{F}_k(t)$, it is called sub-distribution hazard.

The sub-distribution hazard became very popular in recent years since it has direct relationship with CIF and different methods focusing on the sub-distribution hazard have been proposed in regression model (Fine and Gray, 1999) which is discussed later in sections 4.2.1 and 4.2.2.

2.2.4 Relationship Between Cause-specific and Sub-distribution Hazard Rate

The relationship between the cause-specific hazard rate and the sub-distribution hazard rate can be derived analytically via the relationships to the CIF shown in (9), (11) and (12) (Beyersmann and Schumacher, 2007). A detailed derivation of that relationship is presented (Beyersmann et al., 2012). In the case of two possible endpoints,

$$\lambda_1(t) = \gamma_1(t)\left(1 + \frac{\bar{F}_2(t)}{S_{osf}(t)}\right) \tag{13}$$

with $\lambda_1(t)$ denoting the cause-specific hazard for the event of interest ($k = 1$), $\gamma_1(t)$ the corresponding sub-distribution hazard, $\bar{F}_k(t)$ the CIF for the competing event ($k = 2$) and $S_{osf}(t)$ the overall survivor function, providing the probability of freedom from any event up to t time. The relationship given by (13) indicates that the sub-distribution hazard for event type $k = 1$ is related to the cause-specific hazards of both event types, as the cumulative incidence function for event type $k = 2$ and the overall survivor function depend on the cause-specific hazards for both types of event. Therefore, analysis of the cause-specific and the sub-distribution hazards will generally lead to different results in presence of competing risks. Figure 1 depicts the cause-specific and the sub-distribution hazard for an event of interest for various values of the cause-specific hazard for the competing event. For all scenarios, $\lambda_1 = 0.10$ are chosen as the cause-specific hazard for the event of interest and for the competing event, cause-specific hazard (λ_2) are chosen as 0.01, 0.05, 0.10 and 0.25 respectively.

The difference between cause-specific and sub-distribution hazard depends on the risk for a competing event which is driven by the cause-specific hazard $\lambda_2(t)$ (shown in Figure 1). It follows from (13) and from definition of the risk set provided in (10) that the cause-specific and the sub-distribution hazard are equal in absence of competing risks, i.e., in the standard survival setting with one possible endpoint, and that they have to approach the same value for t going to zero whenever competing risks are present. From (11), it follows that the sub-distribution hazard has to converge to zero whenever t tends to infinity, since the CIF approaches a value smaller than one in presence of competing risks, and therefore the cumulative sub-distribution hazard function has to converge to a finite value.

In general, suppose we have m different types of failure, and the respective times to failure are $T_1, T_2, T_3, \ldots, T_m$. But we observe only $T = \min(T_1, T_2, \ldots, T_m)$. Sometimes these $T_1, T_2, T_3, \ldots, T_m$ are called the latent variables. Central to competing risks data is the concept of cause-specific hazard functions, which focuses on what the observed survival is due to a certain cause of failure, while acknowledging that there are other types of failures operating at the same time. There might also be independent censoring C, in which case, we observe $X = \min(T, C)$ and $\delta = I(T < C)$; whereas, $\delta = 0$ is chosen if the case is censoring and various aspects are to be chosen for the purpose of more than one variable.

3 Analysis

Several regression methods in case of competing risks are introduced in recent years. The most commonly used approaches are the cause-specific hazards regression introduced by Prentice et al. (1978) and also the sub-distribution hazards regression proposed by Fine and Gray (1999). In this section, these regression approaches have been described.

3.1 Regression Approaches

3.1.1 Cause-specific Hazards Regression

In survival analysis, the competing risks setting can be used whenever the presence of censored observation has to be considered and we have to estimate the effect of covariates on the cause-specific hazard rates (Prentice et al., 1978). For each individual i the

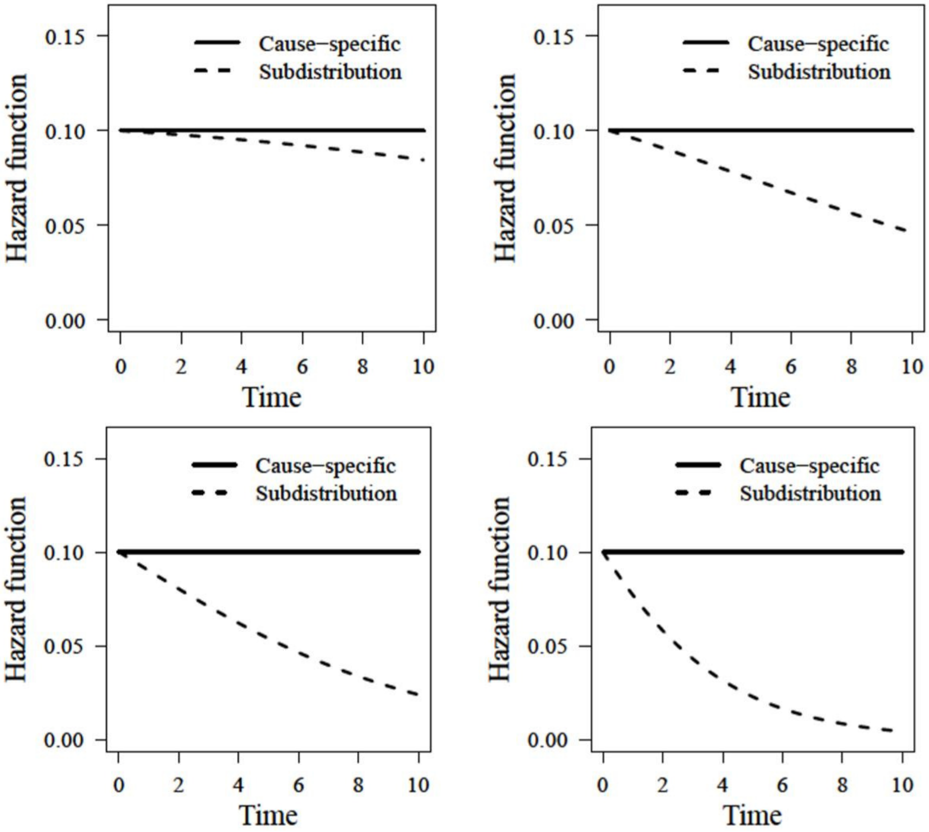

Figure 1: For all scenarios in the four graphs, $\lambda_1 = 0.10$ is chosen as the cause-specific hazard for the event of interest but for the competing event, cause-specific hazard (λ_2) are chosen as 0.01 (top left), 0.05 (top right), 0.10 (bottom left) and 0.25 (bottom right).

data $(t_i, j_i, \delta_i, x_i)$ are observed, where t_i is the observed time, j_i is the observed cause of failure, δ_i is a censoring indicator returning the value of zero for a censored observation and a value of one if any event was observed, and x_i is the vector of covariates, which is assumed to be constant over time and $S(t_i \mid x_i)$ be the conditional survivor function. For a censored observation an arbitrary value can be set for j_i. The likelihood function under independent censoring can be written in the following form:

$$\begin{aligned} L &= \prod_{i=1}^{n} \left(\lambda_{ji}(t_i \mid x_i)^{\delta_i} S(t_i \mid x_i)\right) \\ &= \prod_{i=1}^{n} \left\{\lambda_{ji}(t_i \mid x_i)^{\delta_i} \prod_{l=1}^{K} \exp\left(-\int_0^{t_i} \lambda_l(s \mid x_i)\, ds\right)\right\} \end{aligned} \tag{14}$$

which is an adaptation of the likelihood function used in standard survival analysis considering the relationship between the overall survivor function and the cause-specific hazards shown in (8).

By virtue of the representation of competing risks data and covariates as a triple $(t_i,\ d_i,\ x_i)$ with d_i indicating the type of event ($d_i \in \{1, \ldots, K\}$), or, a censored observation ($d_i = 0$), the likelihood function can be expressed equivalently in the following form:

$$\begin{aligned} L &= \prod_{i=1}^{n} \left(\prod_{l=1}^{K} \left(\lambda_l(t_i \mid x_i)^{I(d_i=l)} S(t_i \mid x_i)\right)\right) \\ &= \prod_{i=1}^{n} \left[\prod_{l=1}^{K} \left\{\lambda_l(t_i \mid x_i)^{I(d_i=l)} \prod_{l=1}^{K} \exp\left(-\int_0^{t_i} \lambda_l(s \mid x_i)\, ds\right)\right\}\right] \end{aligned} \tag{15}$$

The form of the likelihood function presented in Equations 14 or 15 leads to some important implications. These are as follows (Prentice et al., 1978):

(i) The hazard functions and the regression coefficients are identifiable and can be estimated from the observed data.

(ii) The score function for estimation of regression coefficients for the event of interest does not change, when all observed competing events are treated like censored observations. Therefore, standard methods for estimation of hazard rates or hazard ratios can be applied treating competing events as censored observations.

(iii) Covariate effects on cause-specific hazards for different event types can be estimated in separate regression models. When regression models for all event types are fit to the data in order to model the complete competing risks process, different sets of covariates might be considered for different types of event, denoted by an according index.

The effect on the cause-specific hazard of a certain event type does not necessarily translate into an effect on the event probability, represented by the CIF. This fact is further discussed and illustrated in Section 4.1.2, describing how the CIF can be estimated from proportional cause-specific hazards regression models for a given vector of covariates and in Section 4.3 discussing differences between the cause-specific and the sub-distribution hazards regression model.

To estimate covariate effects on the cause-specific hazard rates (Prentice et al., 1978), assuming proportional cause-specific hazards such as

$$\lambda_k(t|x) = \lambda_{k;0}(t)\exp(\beta_k^T x). \tag{16}$$

Here $\lambda_{k;0}(t)$ describes the cause-specific baseline hazard for event type k, which is considered as high-dimensional nuisance parameter, when covariate effects are estimated, x is the P-dimensional vector of covariates and β_k is the vector of regression coefficients of length P for the k^{th} type of event.

Predicting the CIF The CIF for a certain type of event can be estimated under consideration of the covariate information. Assuming the vector of event times with an observed event of type k, denoted as $\underline{t}_k = (\underline{t}_{k1}, \dots, \underline{t}_{kN_k})$ to be ordered and the estimator for the CIF of event type k can be written in the following form:

$$\begin{aligned}\hat{\hat{F}}_k(t|x) &= \sum_{i:\underline{t}_{ki}\le t} \hat{\lambda}_k(\underline{t}_{ki}|x)\hat{S}(\underline{t}_{k(i-1)}|x)\\ &= \sum_{i:\underline{t}_{ki}\le t} \hat{\lambda}_{k;0}(\underline{t}_{ki})\exp(\hat{\beta}_k^T x)\exp\left(-\sum_{l=1}^{K}\hat{\Lambda}_l(\underline{t}_{k(i-1)}|x)\right)\\ &= \sum_{i:\underline{t}_{ki}\le t} \hat{\lambda}_{k;0}(\underline{t}_{ki})\exp(\hat{\beta}_k^T x)\exp\left(-\sum_{l=1}^{K}\hat{\Lambda}_{l;0}(\underline{t}_{k(i-1)})\exp(\hat{\beta}_l^T x)\right)\end{aligned} \tag{17}$$

While competing events can be treated like censored observations for the estimation of cause-specific hazard rates, competing events have to be considered for the estimation of cumulative incidence functions. As can be seen in (17), the cumulative incidence function for event type k depends on the cause-specific hazards of all event types, as previously discussed in Section 3.2. Therefore, an observed effect on the cause-specific hazard does not necessarily translate into an effect on the CIF. This is further discussed in Section 4.3.

3.2 Sub-distribution Hazards Regression

In this section, sub-distribution hazard rate is described to develop a regression model for time-to-event data whenever the competing risks is present. Fine and Gray (1999) proposed to use a Cox regression approach for the sub-distribution hazard for an event of interest, here $k = 1$, assuming proportional sub-distribution hazard rates as follows:

$$\gamma_1(t|x) = \gamma_{1;0}(t)\exp(\eta_1^T x) \tag{18}$$

where $\gamma_1(t|x)$ denotes the sub-distribution hazard for the event of interest depending on the vector of covariates x, $\gamma_{1;0}(t)$ is the baseline sub-distribution hazard for an individual with all covariates equalling zero and η_1 is the vector of regression coefficients. As the competing events are incorporated implicitly in the adapted risk set (Section 3.3) only a model for the event of interest $k = 1$ is presented. In general, the proportionality assumption cannot hold true for separate sub-distribution hazards regression models for different types of event (Beyersmann et al., 2012). Grambauer et al. (2010) investigated the impact of model misspecification. They demonstrated that a sub-distribution hazards regression model has a proper interpretation, even when the sub-distribution hazards were falsely assumed to be proportional. The estimated regression coefficients can be interpreted as average sub-distribution log-hazard ratios. In this case, the average sub-distribution hazard ratio will depend on the length of follow-up (Schemper et al., 2009).

For estimation of the regression coefficients in a sub-distribution hazards regression model a different risk set is needed than for to the cause-specific hazards regression model described in Section 4.1.1. While estimation of the regression coefficients is straightforward when complete data are observed for all individuals and under administrative censoring, the estimating procedure becomes more complicated for incomplete data with non-administrative censoring, in order to obtain unbiased estimates (Fine and Gray, 1999).

Predicting the CIF The predicted CIF for a given vector of covariates x can be obtained from the estimated regression coefficients using the relationship between the sub-distribution hazard and the cumulative incidence function (Equation (11)) without further consideration of effects on the competing events:

$$\begin{aligned}\hat{\hat{F}}_1(t|x) &= 1 - \exp\Big(-\hat{\Gamma}_1(t|x)\Big)\\ &= 1 - \exp\Big(-\int_0^t \hat{\gamma}_1(s|x)\,ds\Big)\\ &= 1 - \exp\Big(-\int_0^t \hat{\gamma}_{1;0}(s)\exp(\hat{\eta}_1^T x)\,ds\Big) \qquad (19)\end{aligned}$$

Estimation of a confidence band for the CIF derived from a proportional sub-distribution hazards model is also important role in regression purpose (Fine and Gray, 1999).

3.3 Differences Between Cause-specific and Sub-distribution Hazards Regression

The two hazard-based regression approaches: (i) the cause-specific and (ii) the sub-distribution are the most popular methods for analysis of competing risks data in medical settings. Due to the similarity of the approaches, the regression coefficients obtained from the regression models are often interpreted in an equal manner without considering that the methods focus on different quantities, namely either the cause-specific or the sub-distribution hazard. Depending on the amount of competing events and on the covariate effects on the competing events, the two approaches can able to provide substantially different regression coefficients since the cause-specific hazards regression aims on the instantaneous risk, whereas the sub-distribution hazard is directly linked to the cumulative incidence function. These differences are displayed and discussed with some simulated examples. Other illustrations can be found in Putter et al. (2007); Allignol et al. (2011); Dignam et al. (2012).

For each scenario competing risks data with two possible endpoints, one event of interest ($k = 1$) and one competing event ($k = 2$), with cause-specific hazards depending on one binary covariate with groups called $P(X = 0)$ and $Q(X = 1)$ are generated for 10,000 subjects. Time-constant cause-specific hazard rates are defined for both groups. So, the assumption of proportionality holds for the cause-specific hazards, leading to time-independent cause-specific hazard ratios. For convenience, only administrative censoring after five years is considered in the examples given in Section 5. Numbers of patients at risk are displayed under the corresponding figures for both groups to illustrate the influence of competing events on the risk set. The cause-specific hazard ratio and the sub-distribution hazard ratio will generally be different and proportionality for one of these measures contradicts proportionality for the other one. For analysis of the simulated data, proportional hazards regression models for the cause-specific and the sub-distribution hazards as described in Sections 4.1.1, 4.1.2, 4.2.1 and 4.2.2 are applied, although the assumption of proportionality is violated for the sub-distribution hazards model. The estimated sub-distribution hazard ratio can be interpreted as average sub-distribution hazard ratio (Grambauer et al., 2010; Hjort, 1992).

3.4 Regression Models Based on Pseudo-observations

By using pseudo-observations, a method for the estimation of covariate effects on state probabilities in multi-state models has been used in various fields (Andersen and Keiding, 2012). Since a competing risks model can be interpreted as a special case of a multi-state model, this approach can be adjusted for the competing risks purpose (Klein et al., 2008). The pseudo-value approach with a complementary log-log link provides results similar to the hazard-based regression models. Generally, the pseudo-observation approach can be considered to estimate effects of covariates on any function of event times $f(T)$, if an unbiased estimator $\hat{\theta}$ exists for

$$\theta = E(f(t)). \qquad (20)$$

A summary of different approaches for survival analysis based on pseudo-observations has been described by Andersen and Perme (Andersen and Perme, 2010). Main idea of the approach is to obtain quantities which allow application of standard methods for data analysis without consideration of censored observations. The estimated pseudo-observations $\hat{\theta}_i$ (where $i = \{1, 2, \ldots, n\}$) which are assessed via leave-one-out estimates (Miller, 1974) for some measure of interest, can be used for that purpose such as

$$\hat{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}^{(i)}. \qquad (21)$$

Here $\hat{\theta}$ is the estimated measure of interest using all n observations and $\hat{\theta}^{(i)}$ points out the estimated measure of interest derived from all but the i^{th} observation. The pseudo observations can be estimated for one fixed time point τ_0, or, for a pre-specified number of time points $\tau = (\tau_1, \dots, \tau_H)$. If multiple time points are considered, an $n \times H$ matrix of pseudo observations is obtained. For regression purposes these pseudo-observations $\hat{\theta}_{ih}$ can be used as response variable (Klein and Andersen, 2005) in a generalized linear model which is as follows:

$$g(\theta_{ih}|x_i) = \alpha_h + \beta^T x_i \tag{22}$$

where $g(.)$ is a link function as the logit or the complementary log-log function and x_i is the vector of covariates of subject i. The influence of the covariates on the pseudo-observations which translates to an influence of the covariates on the measure of interest $f(T)$, can be estimated by using adequate methods for generalized linear models (GLM). In the case of multiple time points, the generalized estimation equation (GEE) approach (Liang and Zeger, 1986) is used for estimation and inference to account for repeated measures on the same subjects in order to obtain robust and valid standard errors under independent censoring. Apart from this, there are many various assumptions present for the working covariance matrix used in the GEE model (Klein et al., 2008). Since any relevant effects did not find of the choice of the working covariance on the estimated regression coefficients and standard errors, so the use of an independent working covariance structure is proposed for this method (Klein et al., 2008).

For the competing risks setting the relevant measure $f(T)$ is the CIF for event type k. So, for each individual i a pseudo-observation $\hat{\theta}_{ih}$ is derived for each of the predefined time points in τ by using the CIF estimated from all subjects and the estimate based on all but the i^{th} individual such as

$$\hat{\theta}_{ih} = n\hat{\bar{F}}_k(\tau_h) - (n-1)\hat{\bar{F}}_k^{(i)}(\tau_h). \tag{23}$$

If censoring is absent in the whole dataset, the pseudo-value indicates, whether subject i failed from cause k up to time τ_h and d_i indicating the type of event where $d_i \in \{1, \dots, k\}$ i.e.,

$$\hat{\theta}_{ih} = \begin{cases} 1 & \text{if } t_i \leq \tau_h \text{ and } d_i = k, \\ 0 & \text{elsewhere.} \end{cases}$$

The mean of the pseudo-values for each considered time point equals the estimate of the CIF. In presence of censored observations, pseudo-values can be smaller than zero for individuals still under observation (i.e., for individuals with a censored observation or for individuals that failed from a competing event, or larger than one after an event of interest is observed, with the actual value depending on the observation time and the amount of censoring). When a complementary log-log link is used between the response (the pseudo-values) and the linear predictor, the regression coefficients can be interpreted as sub-distribution log hazard ratios, if all covariates are time-independent (Klein and Andersen, 2005). The model is as follows:

$$\ln(-\ln(\theta_{ih})) = \alpha_h + \beta^T x_i \tag{24}$$

The analysis can be performed using the R with package "geepack" (Højsgaard et al., 2005; Klein and Andersen, 2005) that allows to specify a complementary log-log link between response and linear predictor.

4 Numerical Results

4.1 Theoretical Simulation Scenarios

For evaluating the performances of the proposed estimators, we conduct simulation studies in different scenarios described in the following and also illustrate the differences in results obtained from cause-specific and sub-distribution hazards regression in various scenarios.

Scenario 1 In the first scenario, the difference between the regression coefficients estimated from a proportional cause-specific hazards and a proportional sub-distribution hazards regression model with two possible endpoints, but a group difference only for the event of interest, is investigated. The cause-specific hazards for the event of interest are chosen to be $\lambda_1(t|X = 0) = 0.2$ and $\lambda_1(t|X = 1) = 0.4$, so a cause-specific hazard ratio of 2 is expected for the event of interest. For the competing event $(k = 2)$ the hazard rates are chosen to be equal for both groups $\lambda_2(t|X = 0) = \lambda_2(t|X = 1) = 0.3$ which implies that there is no group effect on the risk for the competing event. The estimated cause-specific hazard ratio for the event of interest is close to 2, namely $\exp(\hat{\beta}_1) = 2.01$, the estimated sub-distribution hazard ratio for the event of interest is $\exp(\hat{\gamma}_1) = 1.81$, which is slightly smaller than the estimated cause-specific hazard ratio due to the different risk sets used. The estimated cumulative incidence functions for both groups are shown in Figure 2. As the cause-specific hazard for the competing event is the same for both groups, the estimated cumulative incidence functions do not cross.

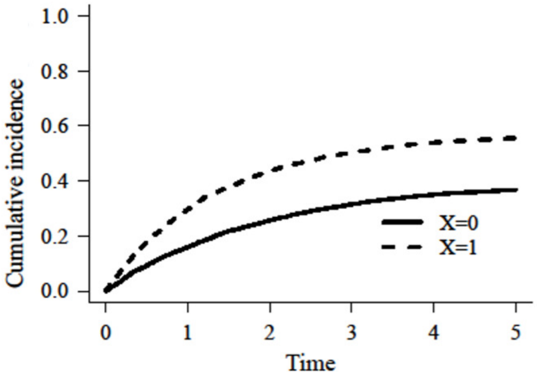

Figure 2: Cumulative incidence function for the event of interest for Scenario 1.

Scenario 2 In the second scenario the cause-specific hazards for the event of interest ($k = 1$) for both groups which gives a cause-specific hazard ratio for the event of interest of $\exp(\beta_1) = HR^{cs}_{k=1} = 2$. The hazard ratio for the competing event ($k = 2$) is defined to be even larger with the cause-specific hazard in group Q being 0.8 and the hazard for group P being 0.2, translating a cause-specific hazard ratio for the competing event of $\exp(\beta_2) = HR^{cs}_{k=2} = 4$. That scenario corresponds to an illustration presented by Putter et al. (2007). The cumulative incidence functions for event $k = 1$ are displayed in Figure 3. Due to the higher amount of competing events in group Q ($X = 1$) compared to group P ($X = 0$), the number of patients at risk is decreasing more slowly in group P. Therefore, a higher incidence of events of interest is observed in group P, although patients of group Q had a higher cause-specific hazard for experiencing an event of type $k = 1$.

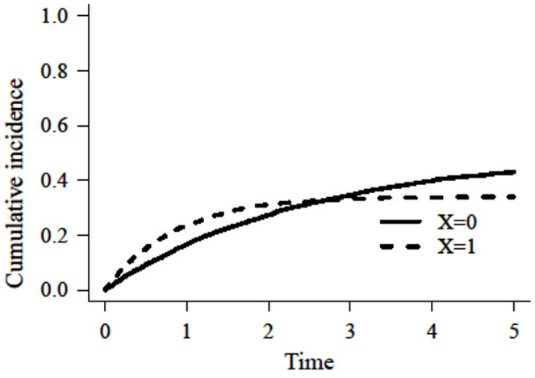

Figure 3: Cumulative incidence function for the event of interest for Scenario 2.

In that situation the higher cause-specific hazard of group Q compared to group P does not translate into a higher incidence of events of type 1 in group Q for late time points. Analysis of the simulated data gives an estimated cause-specific hazard ratio of 1.99, but a sub-distribution hazard ratio of 0.82, revealing different signs of the regression coefficients. The covariate effect on the sub-distribution hazard has to be interpreted as time-averaged effect, since the assumption of proportional sub-distribution hazards is violated. In the sub-distribution hazards regression model, regression coefficients are directly linked to the CIF. Since the sub-distribution hazard for the event of interest is higher for group P than for group Q for most time points, a higher average sub-distribution hazard for group P is estimated, translating to an average sub-distribution hazard ratio smaller than one. Cause-specific hazards regression shows the covariate effect on the instantaneous risks, and sub-distribution hazards regression represents the effect on the cumulative incidence function which lead to various fruitful conclusions regarding the covariate effect on the event of interest.

Scenario 3 In a third scenario the setting is similar to scenario 2, but with a much lower cause-specific baseline hazard for the competing event ($\lambda_2(t|X = 0) = 0.05$, $\lambda_2(t|X = 1) = 0.2$), leading to a smaller amount of observed events of type $k = 2$. In the simulations, 6029 events of interest are observed but only 2297 individuals fail from a competing event. In this case, the difference between estimated cause-specific and sub-distribution hazard ratios is smaller than in scenario 2 with $\exp(\hat{\beta}_1) = 1.95$ and $\exp(\hat{\gamma}_1) = 1.28$. The estimated cumulative incidence functions obtained from the simulated dataset are shown in Figure 4.

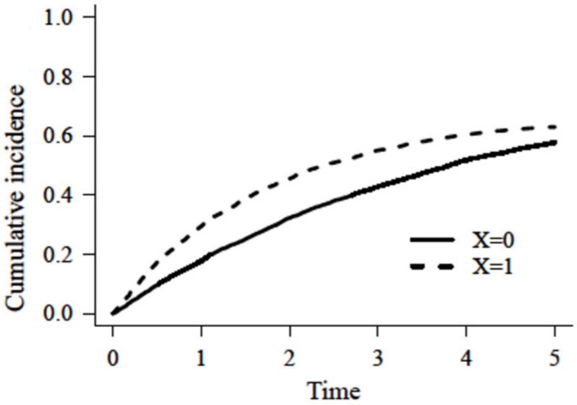

Figure 4: Cumulative incidence function for the event of interest for Scenario 3.

Scenario 4 The cause-specific hazard rates for the event of interest are chosen to be equal for both groups, leading to a cause-specific hazard ratio of one ($\lambda_1(t|X = 0) = 0.4$, ($\lambda_1(t|X = 1) = 0.4$, $\exp(\beta_1) = HR^{cs}_{k=1} = 1$). For the competing event, a cause-specific hazard ratio of $\exp(\beta_2) = 3$ is chosen for the simulation ($\lambda_2(t|X = 0) = 0.1$, $\lambda_1(t|X = 1) = 0.3$). The corresponding cumulative incidence functions are displayed in Figure 5.

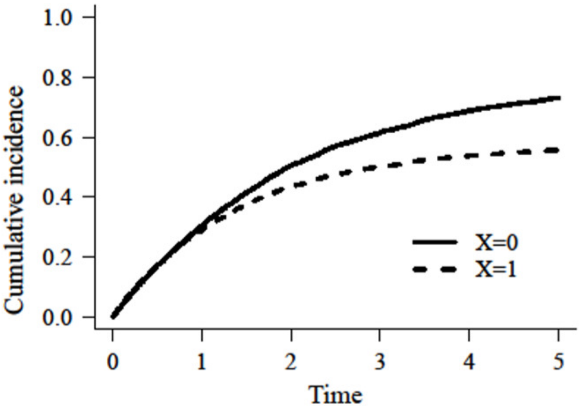

Figure 5: Cumulative incidence function for the event of interest for Scenario 4.

Due to the different risks for the competing event, leading to a higher number of competing events in group Q than in group P, the number of patients at risk decreases faster in group Q. Therefore, a higher incidence of events of interest is observed in group P compared to group Q. A cause-specific hazard ratio of 1.03 is estimated, whereas sub-distribution hazards regression reveals a hazard ratio of 0.70, since the cumulative incidence curves differ between both groups. In such a situation, careless interpretation of the sub-distribution hazards regression coefficient might lead to biological implausible conclusions, interpretation of the cause-specific hazards regression coefficient for the event of interest, ignoring the effect on the competing event, will miss important information on the group difference regarding the other type of event and consequently on the event probabilities for the event of interest.

The simulations described here reveal that substantial differences in the results of cause-specific and sub-distribution hazards regression may be present in certain scenarios. Careless interpretation of the estimated regression coefficients may lead to wrong conclusions regarding associations between covariates and risks or event probabilities. Therefore, investigators should be aware of differences between cause-specific hazards and sub-distribution hazards regression to avoid misuse of the methods and misinterpretation of obtained results. Both regression models are applied to a real data example for investigation of occurrence of blood stream infection during neutropenia (when a person has neutrophils, i.e., an abnormally low count of a type of white blood cell) after peripheral blood stem-cell transplantation and also compared and discussed differences in the methods and in the obtained results (Beyersmann et al., 2007). Besides, Latouche et al. (2013) have recommended to present covariate effects obtained from cause-specific hazards regression models for all possible types of event and from a sub-distribution hazards regression model for the event of interest, accompanied by estimates of the cumulative incidence functions, to assess whether there is

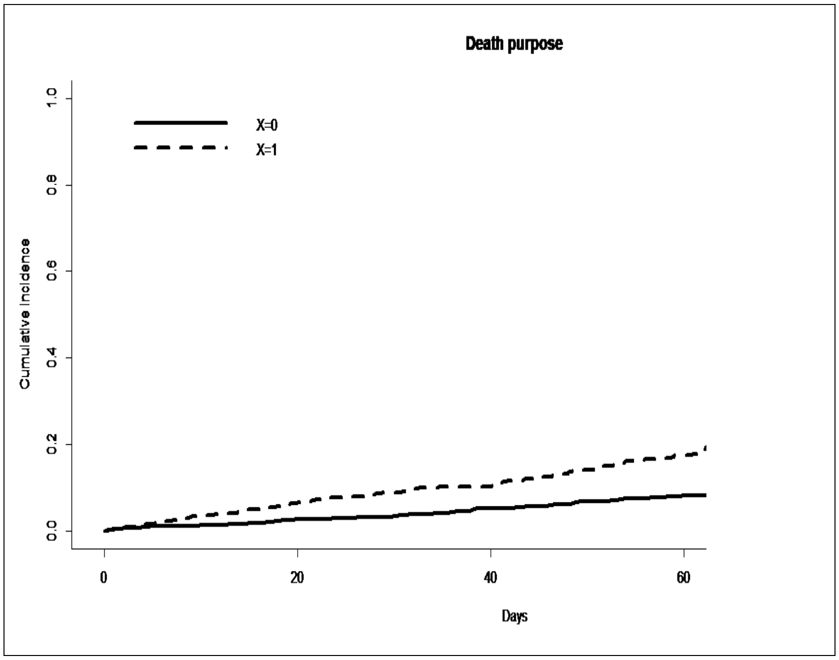

Figure 6: Cause-specific hazards regression model: $X = 0 \longrightarrow$ Italy and $X = 1 \longrightarrow$ USA.

a direct effect of the covariate of interest on the CIF (as in scenario 1) or an indirect effect caused by an effect on the competing event(s) (as in Scenario 4). Presentation of results obtained from the different regression models and display of the cumulative incidence functions should avoid pitfalls and possible misinterpretations discussed in the examples here.

4.2 Application of Regression Approaches to COVID-19 Outbreak in USA

Example 1 In this section, we have described substantial differences in the results of cause-specific and sub-distribution hazards regression dataset of the patients due to SARS-CoV2 or COVID-19. In this work, competing risks data with two possible endpoints, one event of interest ($k = 1$) and one competing event ($k = 2$), with cause-specific hazards depending on one binary covariate with groups called P ($X = 0$) and Q ($X = 1$) are generated for COVID-19 patients. In this present study, we will focus on an example studying patient survival on COVID-19, where death is the event of interest. So, in this work we proceed with the mortality rate and illustrate only the prediction related to the death with COVID-19. Firstly, our aim is to demonstrate the impact of considering competing risks to estimate the cumulative incidence function. We compare the event of interest (i.e., mortality rate) among USA and Italy choosing as two groups in this work. It is assumed that P ($X = 0$) corresponds to Italy and Q ($X = 1$) corresponds to USA. Figure 6 demonstrates that the death rate is highly increasing in USA than Italy during this prevalence of coronavirus disease. In survival analysis, conventional methods ignore the competing events such as the Kaplan Meier (KM) method, standard Cox proportional hazards regression (Noordzij et al., 2013).

As this leads to a difference in the estimated cumulative incidence functions for $k = 1$, which are shown in Figure 6, that is larger than in the absence of competing events, the estimated sub-distribution hazard ratio is larger than the estimated cause-specific hazard ratio with $\exp(\hat{\gamma}_1) = 1.84$ and $\exp(\hat{\beta}_1) = 1.13$. Due to the opposite direction of the cause-specific hazard ratios for both event types, leading to the same overall hazard, which is defined as the sum of the cause-specific hazards for both types of event as described in (6), the number of patients at risk are similar in both groups for all considered time points.

Example 2 We extend our results to regression analyses that allow to investigate necessary explanatory variables, which are topics currently being pursued, using standard regression models for competing risks data of COVID-19 patients. Application of competing risks regression models including details on the applied methods and results obtained from analyses which are presented in this section. The effect of risk group allocation on cause-specific hazards adjusted for age, gender and also a proportional sub-distribution hazards model fit to the data set of prevalence due to COVID-19 in the New York city where is being hit the hardest by the novel coronavirus spreading across the U.S. In this work, the effects of response variables, i.e., gender (1 as male, 2 as female) and age (0 as < 65, 1 as ≥ 65) are analyzed. The exploratory analysis of COVID-19 data set of New York city is shown in Table 1.

The *R* function "*coxph*" from library "*survival*" is used for estimation of the regression coefficients discussed in section 4.1.1

Table 1: Exploratory analysis of COVID-19 data (consider mortality rate of New York).

Variable	Description	Statistical Summary	Percentage
Gender	Gender	1 ≡ Male (10961) 2 ≡ Female (6721)	62% 38%
Age	Age of patient (in years)	0 ≡< 65 (4954) 1 ≡≥ 65 (12728)	28% 72%
Time	Reporting date (in days)	60	–

(Therneau, 2011). The incidence data set of New York city due to COVID-19 are shown in both (gender and age) purpose in Figures 7 and 8. The bootstrap method is used to quantify the uncertainty associated with a given statistical estimator or with a predictive model. It consists of randomly selecting a sample of n observations from the original data set. This subset, called bootstrap data set is then used to evaluate the model (Lau et al., 2009). So, at first we have demonstrated the coefficient of determination (R^2) for the models, confidence intervals for an R^2 value (proportion of variation in the outcome explained by the predictor variables included in the model) using bootstrap method for goodness-of-fit of the models and model validation purpose. In both cases R^2 is greater than 0.9 (i.e., 0.994 with 95% C.I. (0.853, 1.132) and 0.996 with 95% C.I. (0.856, 1.137) for cause-specific and sub-distribution regression model respectively) which signifies for better fit of underlying models.

Both gender and age have a significant effect on the cause-specific hazards for this type of event (results are shown in Table 2). The result provides more than 2 times higher risk of dying from a COVID-19 for patients who are older than 65 years and the risk is also higher nearly 2 times more if the person being male. Because from Table 2 we get exp(co.eff.) = 1.824 which indicates that the risk is higher (nearly 2 times more) if the person being male. We also get exp(co.eff.) = 2.223 (from Table 2) which suggests that the risk of dying is higher (more than 2 times) if the patients are more than 65 years old.

Table 2: Result of the cause-specific hazards regression model.

	co. eff.	exp(co. eff.)	Std. error	P-value
Gender	0.601	1.824	0.337	<0.01
Age	0.800	2.223	0.375	<0.01

Cumulative incidence functions are predicted from the Cox regression models following (17) in section 4.1.2 using the mean of gender, i.e., the proportion of patients with gender male (62.5%), and the mean of the indicator variable for age, i.e., the proportion of patients being at least 65 years of age (75%). Cause-specific baseline hazards, which are required for calculation of cumulative incidence functions are derived and the predicted cumulative incidence curves are displayed in Figure 9.

A proportional sub-distribution hazards model as described in Equation (18) is fit to the data in order to assess the influence of gender and age on the sub-distribution hazards for both types of event. The analysis is performed using the function "*crr*" in the *R* library "*cmprsk*" and the results from the sub-distributional hazards models have to be interpreted as time-averaged effects (Grambauer et al., 2010).

Results of the regression model investigate the influence of covariates on the sub-distributional hazards and provide that both gender and age have a significant effect on the these underlying hazards same as cause-specific hazards for this type of event (shown in Table 3). Effects on the sub-distributional hazards can be translated directly to effects on the cumulative incidence functions and the predicted cumulative incidence curves are displayed in Figure 10. In order to analyse the data using the approach which is sketched in section 4.4, pseudo values are estimated for each individual and used as response in a GEE model (Klein et al., 2008). Besides, Andersen et al. (Andersen et al., 2003) introduced a calculation technique for estimation of covariate effects on event probabilities in multi-state models using pseudo-values, that are derived by jackknife estimates from the original data. In a first step, the CIF for death due to COVID-19 is estimated for 4 different points in time (2 weeks intervals equally spaced from baseline to 2 months of follow-up) for the whole data-set. Pseudo-observations are calculated from these 17,671 × 4 estimates

Table 3: Result of the sub-distribution hazards regression model.

	co. eff.	exp(co. eff.)	Std. error	P-value
Gender	0.601	1.824	0.327	<0.01
Age	0.800	2.225	0.291	<0.01

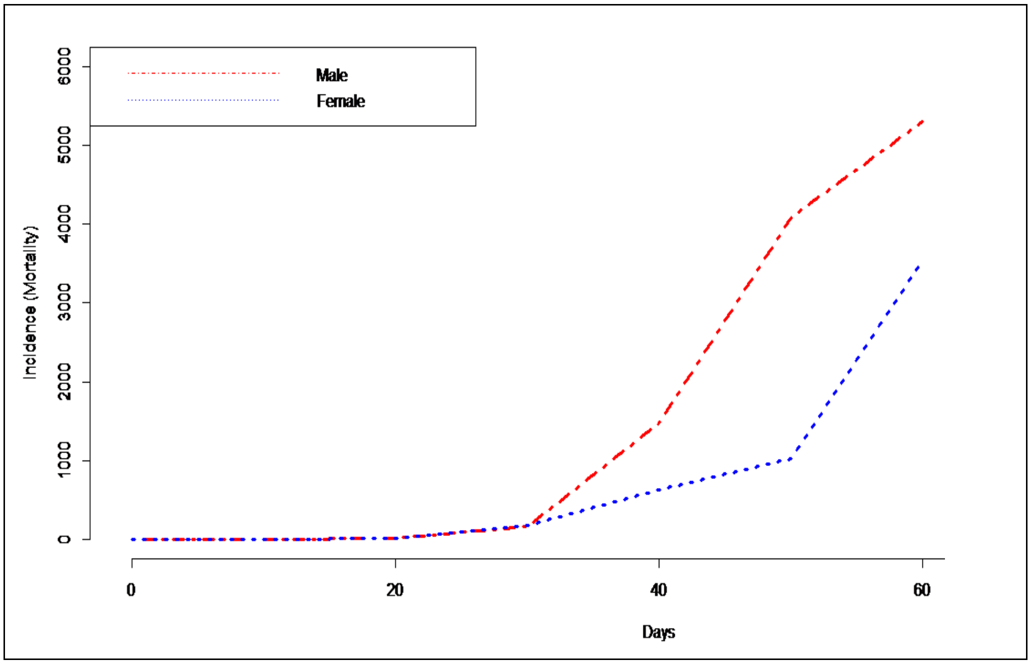

Figure 7: Incidence data (mortality rate) for gender purpose in New York.

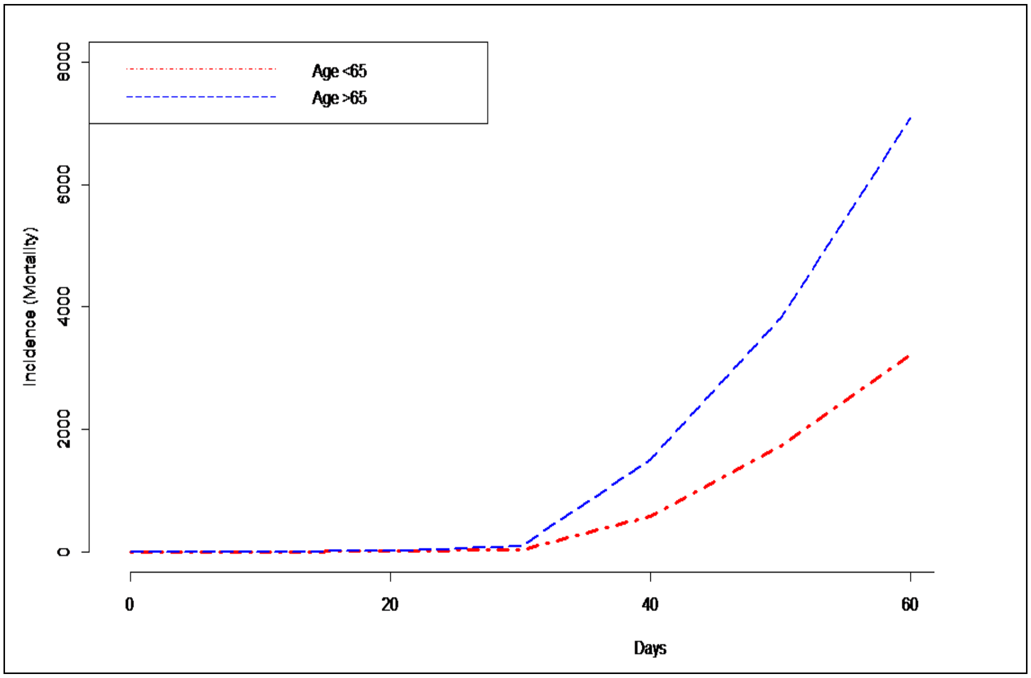

Figure 8: Incidence data (mortality rate) for age purpose in New York.

Table 4: Regression coefficients obtained by the pseudo-value approach.

	co. eff.	exp(co. eff.)	Std. error	P-value
Constant	2.317	10.145	0.730	<0.01
Gender	0.681	1.976	0.408	<0.01
Age	0.855	2.351	0.360	<0.01
Time = 2 weeks	1.493	4.450	0.299	<0.01
Time = 4 weeks	2.199	9.016	0.331	<0.01
Time = 6 weeks	2.596	13.410	0.353	<0.01
Time = 8 weeks	4.008	54.652	0.989	<0.01

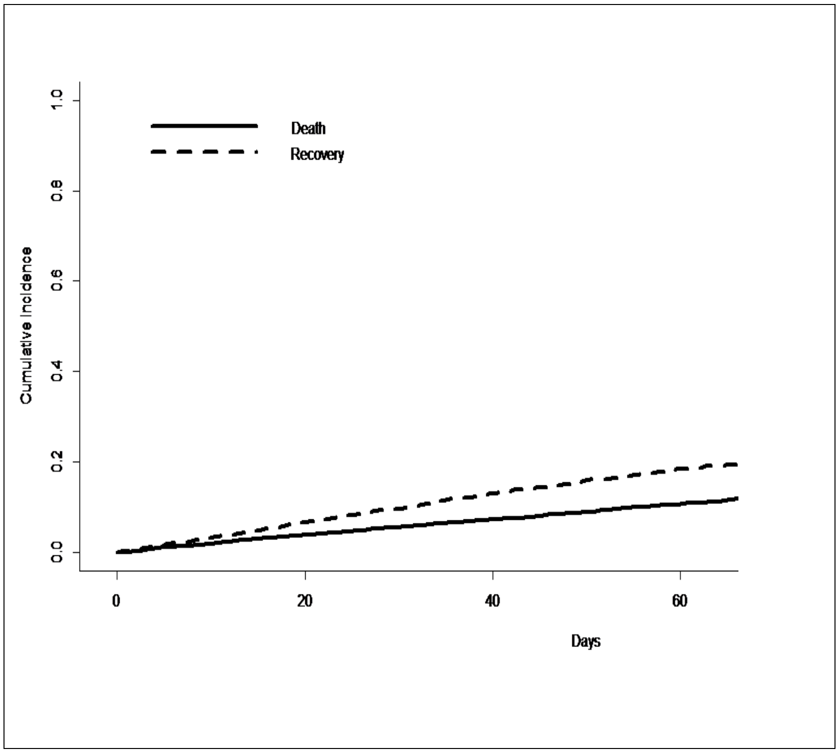

Figure 9: Estimated CIF for COVID-19 for cause-specific hazards regression.

following Equation (23).

In this analysis, the estimation of the cumulative incidence functions, are monotonously increasing and $\exp(\hat{\beta})$ for gender and age can be interpreted as sub-distribution hazard ratio shown in Table 4. These pseudo-values are used as dependent variables in a GEE model, to account for multiple observations of the same subjects and 4 dummy variables indicating the time point are included as covariates. The independent working covariance matrix is used in the GEE model (Klein et al., 2008).

The influence of the covariates of interest on the pseudo-values is estimated using a complementary log-log (cloglog) link between the response and the linear predictor, applying the function "*geese*" of the *R* library "*geepack*" (Højsgaard et al., 2005). So, the estimated coefficients can be interpreted as logarithms of sub-distribution hazard ratios. The results of the GEE model are presented in Table 4. Effects observed in the pseudo-value approach are similar to those obtained in the Fine and Gray model (Fine and Gray, 1999) and can be interpreted analogously as effects on the sub-distribution hazard, translating to effects on the CIF (Klein et al., 2008). As described by Andersen and Perme (2010), the standard errors obtained in the pseudo-value approach are higher than those in the Fine and Gray regression model (Fine and Gray, 1999). Regression coefficients for the different time points specified for calculation of the pseudo observations, which are partly presented in Table 4, are not of major interest, but are necessary for estimation of the CIF. The estimated CIF derived from results of the pseudo-observation approach, which is shown in Figure 11, is similar to the cumulative incidence functions obtained from the cause-specific hazards regression or the sub-distributional hazards regression. The logistic regression ensuring the determination of the risk factors as probability is a method that investigates the relationship of the result variables with independent variables in binary or multiple phases in all

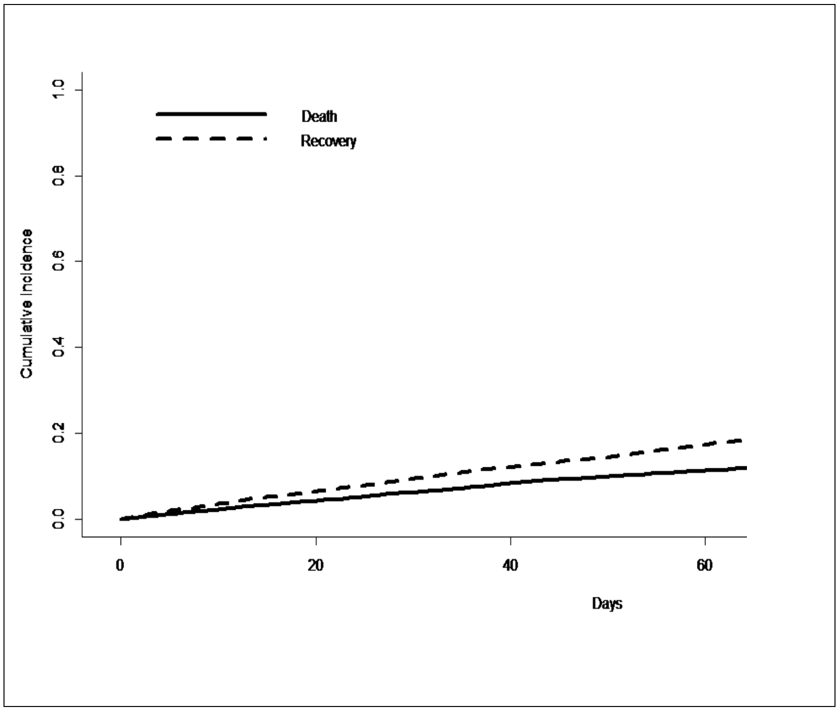

Figure 10: Estimated CIF for COVID-19 for sub-distribution hazards regression.

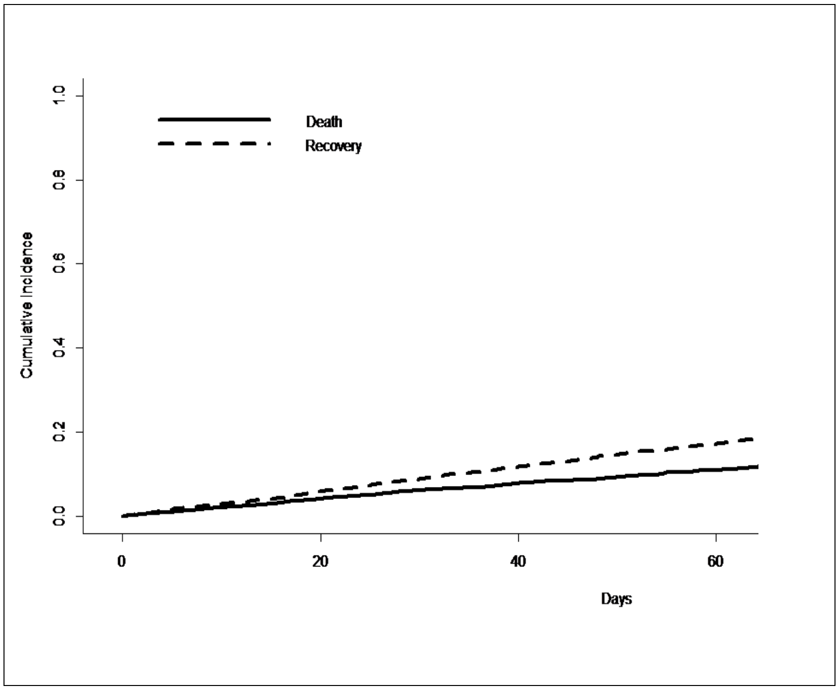

Figure 11: Estimated CIF for COVID-19 for the analysis based on pseudo observation.

Table 5: Comparison of models with AIC and BIC.

Model	AIC	BIC
Cause-specific	300.497	304.361
Sub-distributional	300.001	304.051

areas of public health research (Agresti, 2007; Ghosh and Samanta, 2019a,b,c). So, it is also demonstrated how the coefficients can be interpreted after using logistic regression (also known as logit model) for the underlying dataset. The GLM logit model provides that more than 1.23 times higher risk of dying from a COVID-19 for patients who are older than 65 years and the risk is also higher near about 0.92 times more if the person being female.

4.3 Comparison of Models

One of the problems with the implementation of GEE models is that GEE is a non–likelihood-based method. Therefore, information criteria such as: (i) Akaike Information Criterion (AIC) and (ii) Bayesian Information Criterion (BIC) cannot be directly applied, which creates problems with the choice of best model. So, in this work we have evaluated the Akaike selection criterion (Akaike, 1974) and Bayesian Information Criterion, used to choose only between cause-specific and sub-distributional hazard models (Kuk and Varadhan, 2013).

Table 5 provides the AIC and BIC value of the underlying models which measure goodness of fit. Generally, a good model is the one that has minimum AIC and BIC among all the other models. A lower AIC or BIC value indicates a better fit. So, sub-distributional model (AIC 300.001 and BIC 304.051 as shown in Table 5) is better than cause-specific model (AIC 300.497 and BIC 304.361 as shown in Table 5) for the underlying dataset. The application of the pseudo-value approach with a complementary log-log link does not have any advantages over the sub-distribution hazards regression since it leads to similar results with larger standard errors, which could also be observed for this data-set shown in Table 3 and Table 4. So, overall it can be concluded that sub-distribution hazards regression is best among all underlying models.

5 Discussion

Adequate analysis of competing risks data is relevant for various applications. In medical research, time to a certain cause of death might be of major interest in order to assess efficacy of a therapy or the predictive or prognostic effect of a certain risk factor, with other causes of death being competing risks. This work describes different methods for analysis of competing risks data. The availability of methods for adequate analysis of competing risks data have been assessed and the application of competing risks methods for analysis and presentation of clinical data have also been investigated (Koller et al., 2012).

5.1 Advantages of Underlying Models

The adequate choice of the methods to use is still under discussion, but in recent years most authors are arguing for modeling the whole competing risks process, which is naturally defined by the cause-specific hazard rates (Beyersmann et al., 2007; Andersen and Keiding, 2012; Koller et al., 2012). The sub-distribution hazards regression allows a direct translation of the covariate effects on the hazard rate to an effect on the event probability, which appears to be much more intuitive for applicants and readers not familiar with the concept of hazard rates. For estimation of the sub-distribution hazard rate in the presence of censored observations, a potential censoring time has to be determined for each individual to obtain unbiased estimates. While use of the sub-distribution hazard rate appears appealing due to its direct relationship to the cumulative incidence function, its use is argued against, because of the unintuitive risk set formulation (Andersen and Keiding, 2012). The application of the pseudo-value approach with a complementary log-log link does not have any advantages over the sub-distribution hazards regression since it leads to similar results with larger standard errors, which could also be observed for this data-set, but the pseudo value approach in general allows more flexible modeling in settings where the proportional hazards assumption does not hold. Another advantage is that, by definition, the CIF of each competing event is a fraction of the $S(t)$, therefore the sum of each individual hazard for all competing events should equal the overall hazard. This property of CIF makes it possible to dissect overall hazard, which has more practical interpretations. Apart from this, another advantage of this cause-specific proportional hazard model is that it is easy to fit (by simply censoring for competing events) with any type of statistical software. It is important to realize, however, that because the competing events are treated as censored observations, during follow-up, the number of patients at risk is reduced (Noordzij et al., 2013). The cause-specific approach is that the estimated HR can be interpreted as an HR among those patients who are alive and did not receive a transplant before. Another advantage of the cause-specific approach is that it is easier to handle time-dependent covariates than with the sub-distribution hazards model (Noordzij et al., 2013). Cause-specific model measures the association of an exposure on the corresponding event in which the competing event contributes only by passively removing individuals from the risk set whereas sub-distributional model measures the association of an exposure to the corresponding event in which the competing event actively contributes to the risk set. But both does not have to correctly specify the unspecified baseline cause-specific hazard function (Lau et al., 2009).

5.2 Concluding Remarks

In the present work, both gender and age have a significant role in the prevalence predicted from the underlying approaches with current coronavirus data. The results provides that more than 2 times higher risk of dying from a COVID-19 for patients who are older than 65 years and the risk is also higher near about 2 times more if the person being male. Due to the high amount of censored observations, results from the two hazard-based methods are similar in this example. One major task remains the transfer of available methods for the analysis of competing risks data to the medical community, in order to avoid misinterpretation of study data, possibly leading to erroneous therapy decisions or risk stratifications, due to inadequate application of statistical methods in the presence of competing risks. In this work, it is also observed that the recovery rate is higher than mortality rate although mortality rate is increasing with the time-points because death rate due to COVID-19 is lower than confirmed cases. According to data and statistics website Worldometer, the total number of 2,653,116 confirmed cases and 185,056 deaths from the coronavirus (COVID-19 outbreak as of April 23) were reported worldwide. Community transmission is evidenced by the inability to relate confirmed cases through chains of transmission for a large number of cases, or by increasing positive tests through sentinel samples. New York city is being hit the hardest by the novel coronavirus spreading across the U.S. The more than 5,600 deaths in the city account for roughly one-third of all confirmed U.S. deaths from COVID-19, the illness caused by the coronavirus reported by WHO (Hawkins et al., 11 Apr. 2020, accessed in April 2020). Interrupt human to human transmission including reducing secondary infections among close contacts and health care workers, preventing transmission amplification events, and preventing further international spread. WHO is providing guidance on early investigations, which is critical in an outbreak of a new virus. The data collected from the protocols can be used to refine recommendations for surveillance and case definitions, to characterize the key epidemiological transmission features of COVID-19, help understand spread, severity, spectrum of disease, impact on the community and to inform operational models for implementation of countermeasures such as case isolation, contact tracing and isolation.

Acknowledgments

The authors are grateful to the anonymous referees, Dr. Olcay Akman, Editor-in-Chief, for their careful reading, valuable comments and helpful suggestions, which have helped them to improve the presentation of this work significantly. The last author (AM) would like to acknowledge partial support from the *Expeditions in Computing: Global Pervasive Computational Epidemiology* (NSF Project Number: 1918656) grant as well as from the *Mathematical Theoretical Biology Institute (MTBI)* program (NSF DMS1757968) grant.

References

Agresti, A. (2007). *An introduction to categorical data analysis (2nd edition)*. New York: John Wiley and Sons. 44

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*(6), 716–723. 44

Allignol, A., M. Schumacher, C. Wanner, C. Drechsler, and J. Beyersmann (2011). Understanding competing risks: a simulation point of view. *BMC medical research methodology 11*, 86. doi: 10.1186/1471-2288-11-86. 35

Andersen, P. K. and N. Keiding (2012). Interpretability and importance of functionals in competing risks and multistate models. *Statistics in Medicine 31*, 1074–1088. doi: 10.1002/sim.4385. 30, 35, 44

Andersen, P. K., J. P. Klein, and S. Rosthøj (2003). Generalised linear models for correlated pseudo observations, with applications to multi-state models. *Biometrika 90*, 15–27. doi: 10.1093/biomet/90.1.15. 40

Andersen, P. K. and M. P. Perme (2010). Pseudo-observations in survival analysis. *Statistical Methods in Medical Research 19*, 71–99. doi: 10.1177/0962280209105020. 35, 42

Austin, P.C., D. S. Lee, and J. P. Fine (2016). Introduction to the Analysis of Survival Data in the Presence of Competing Risks. *Circulation 133*(6), 601–609. https://www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.115.017719. 30

Beyersmann, J., M. Dettenkofer, H. Bertz, and M. Schumacher (2007). A competing risks analysis of bloodstream infection after stem-cell transplantation using sub distribution hazards and cause-specific hazards. *Statistics in Medicine 26*, 5360–5369. doi: 10.1002/sim.3006. 38, 44

Beyersmann, J. and M. Schumacher (2007). Letter to the editor: Misspecified regression model for the subdistribution hazard of a competing risk. *Statistics in Medicine 26*, 1649–1652. doi: 10.1002/sim.2727. 32

Beyersmann, J., M. Schumacher, and A. Allignol (2012). *Competing Risks and Multistate Models with R*. New York: Springer. 30, 32, 34

Dignam, J. J., Q. Zhang, and M. Kocherginsky (2012). The use and interpretation of competing risks regression models. *Clinical Cancer Research 18*(8), 2301–2308. doi: 10.1158/1078-0432.ccr-11-2097. 35

Fine, J. P. and R. J. Gray (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association 94*, 496–509. doi: 10.2307/2670170. 32, 34, 35, 42

Ghosh, S. and G. P. Samanta (2019a). Fitting Cumulative Logit Models for Ordinal Response Variables in Retail Trends and Predictions. *International Journal of Statistics and Economics 20*(1), 32–49. 44

Ghosh, S. and G. P. Samanta (2019b). Model Justification and Stratification for Confounding of Chlamydia Trachomatis Disease. *Letters in Biomathematics 6*(2). doi: 10.1080/23737867.2019.1654418. 44

Ghosh, S. and G. P. Samanta (2019c). Statistical modeling for cancer mortality. *Letters in Biomathematics 6*(2). doi: 10.1080/23737867.2019.1581104. 44

Grambauer, N., M. Schumacher, and J. Beyersmann (2010). Proportional subdistribution hazards modeling offers a summary analysis, even if misspecified. *Statistics in Medicine 29*, 875–884. doi: 10.1002/sim.3786. 34, 35, 40

Gray, R. (1988). A class of k-sample tests for comparing the cumulative incidence function in the presence of a competing risk. *The Annals of Statistics 16*, 1141–1154. `https://www.jstor.org/stable/2241622`. 31

Hawkins, D., M. Iati, H. Knowles, S. Denyer, M. Kornfield, T. Bella, and J. Dougherty (2020). Confirmed U.S. covid-19 death toll reaches 20,000, highest in the world. *The Washington Post*. April 11, 2020. `https://www.washingtonpost.com/world/2020/04/11/coronavirus-latest-news/`. Last accessed in April 2020. 45

Hjort, N. L. (1992). On inference in parametric survival data models. *International Statistical Review*, 60: 355-387. 35

Højsgaard, S., U. Halekoh, and J. Yan (2005). The R package geepack for generalized estimating equations 15:1-11, `http://CRAN.R-project.org/package=survival`, R package version 2.36-5. 36, 42

Kaggle. *Kaggle: Novel Coronavirus 2019 data set*. `https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset`. Last accessed in April 2020. 30

Klein, J., M. Gerster, P. Andersen, S. Tarima, and M. Perme (2008). SAS and R functions to compute pseudo-values for censored data regression. *Computer Methods and Programs in Biomedicine 89*(3), 289–300. doi: 10.1016/j.cmpb.2007.11.017. 35, 36, 40, 42

Klein, J. P. and P. K. Andersen (2005). Regression modeling of competing risks data based on pseudo values of the cumulative incidence function. *Biometrics 61*, 223–229. doi: 10.1111/j.0006-341X.2005.031209.x. 36

Koller, M. T., H. Raatz, E. W. Steyerberg, and M. Wolbers (2012). Competing risks and the clinical community: irrelevance or ignorance. *Statistics in Medicine 31*, 1089–1097. doi: 10.1002/sim.4348. 44

Kuk, D. and R. Varadhan (2013). Model selection in competing risks regression. *Statistics in Medicine 32*, 3077–3088. 44

Lambert, P. C., P. W. Dickman, C. P. Nelson, and P. Royston (2010). Estimating the crude probability of death due to cancer and other causes using relative survival models. *Statistics in Medicine 29*, 885–895. doi: 10.1002/sim.3762. 30

Lam, W. K., N. S. Zhong, and W. C. Tan (2003). Overview on SARS in Asia and the World. *Respirology 8*, S2–S5. 29

Latouche, A., A. Allignol, J. Beyersmann, M. Labopin, and J. P. Fine (2013). A competing risks analysis should report results on all cause-specific hazards and cumulative incidence functions. *Journal of Clinical Epidemiology 66*(6), 648–653. doi: 10.1016/j.jclinepi.2012.09.017. 38

Lau, B., S. R. Cole, and S. J. Gange (2009). Competing Risk Regression Models for Epidemiologic Data. *American Journal of Epidemiology 170*(2), 244–256. 40, 45

Liang, K. Y. and S. L. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika 73*, 13–22. doi: 10.1093/biomet/73.1.13. 36

Miller, R. G. (1974). The jackknife-a review. *Biometrika 61*(1), 1–15. doi: 10.1093/biomet/61.1.1. 35

Noordzij, M., K. Leffondré, K. J. Stralen, C. Zoccali, F. W. Dekker, and K. J. Jager (2013). When do we need competing risks methods for survival analysis in nephrology? *Nephrology Dialysis Transplantation 28*(11), 2670–2677. doi: 10.1093/ndt/gft355. 39, 44

Park, S. E. (2020). Epidemiology, virology, and clinical features of severe acute respiratory syndrome –coronavirus-2 (SARS-CoV-2; Coronavirus Disease-19). *Clin Exp Pediatr*. doi: 10.3345/cep.2020.00493. 29

Pintilie, M. (2007). Analysing and interpreting competing risk data. *Statistics in Medicine 26*, 1360–1367. doi: 10.1002/sim.2655. 30

Prentice, R., J. Kalbfleisch, A. Peterson, N. Flournoy, V. Farewell, and N. Breslow (1978). The analysis of failure times in the presence of competing risks. *Biometrics 34*, 541–554. doi: 10.2307/2530374. 32, 33, 34

Putter, H., M. Fiocco, and R. B. Geskus (2007). Tutorial in biostatistics: Competing risks and multi-state models. *Statistics in Medicine 26*(11), 2389–2430. doi: 10.1002/sim.2712. 35, 37

Resche-Rigon, M. and S. Chevret (2006). Local influence for the subdistribution of a competing risk. *Statistics in Medicine 25*, 1937–1947. doi: 10.1002/sim.2354. 30

Schemper, M., S. Wakounig, and G. Heinze (2009). The estimation of average hazard ratios by weighted cox regression. *Statistics in Medicine 28*, 2473–2489. doi: 10.1002/sim.3623. 34

Therneau, T. (2011). "survival: Survival analysis, including penalised likelihood." `http://CRAN.R-project.org/package=survival`, R package version 2.36-5. 40

Tsiatis, A. A. (2005). Competing risks. *In:* Armitage, P. and T. Colton (eds). *Encyclopedia of Biostatistics, (second edition)*, pp. 824–835, New York: John Wiley and Sons. 30