INFERENTIAL PROCEDURE FOR COMPETING RISKS
CURRENT STATUS DATA WITH CONTINUOUS OBSERVATIONS TIMES

C. Silambarasan and R. Elangovan
Department of Statistics
Annamalai University, Tamil Nadu, India.

Abstract

Competing risk data are encountered when subjects under study are at risk of more than one mutually exclusive event, like death from different causes. The term competing risks also refers to data where the different possible events are not mutually exclusive but the interest lies on the first coming event. Current status data arise naturally in medical studies where the target measurement is the time of event occurrence, but observations are limited to indicators of whether or not the event has occurred at the time the sample is collected. In competing risks set up Jewell et al. (2003) introduced simple parametric models, which use exponential distribution to model cumulative incidence functions. It is proposed to study the inferential procedure for competing risks current status data with continuous observations times. Both parametric and nonparametric estimation for the competing risks current status data are also to be compared with possible illustrations. Following the approaches of Jewell, Maathuis and Hudgens, it is also proposed to derive the new inferential procedures for competing risk current status data. Further, by using several ‘R’ routines competing risks current status data are to be analyzed through real data example.

1. Introduction

Survival analysis is the analysis of data measured from a specific time of origin until an event of interest or a specified endpoint refers to Collett, (1994). For example, in order to determine the incidence of death due to breast cancer among breast cancer patients, every patient will be followed from a baseline date (such as date of diagnosis or date of surgery) until the date of death due to breast cancer or study closing date. A patient who dies of breast cancer during the study period would be considered to have an ‘event’ at their date of death.
Clinical studies often focus on estimating the survivor function or the overall survival probability. This is the probability of being event-free at least up to a given time. The event is any specific event of interest. The overall survival probability is estimated using the person follow-up time and event status. The survival at a given time is the conditional probability of surviving to a specific time given that the individual is at risk for the event (such as mortality) at that time. This is estimated as the ratio of the number of individuals that are event-free at that time to the number of individuals that lived event-free at least up to that time. Hence forth, when referring to survival at a given time, we in fact mean survival conditional upon being at risk at that time.

The Kaplan-Meier method for estimating survival functions and the Cox proportional hazards model for estimating the effects of covariates on the hazard of the occurrence of the event are commonly used statistical methods for the analysis of survival data. Competing risks occur frequently in the analysis of survival data. A competing risk is an event whose occurrence precludes the occurrence of the primary event of interest. Current status data with competing risks arise in cross-sectional studies that assess the current status of individuals in the sample with respect to an event that can be caused by several mechanisms.

Hazard Function Regression

A key concept in survival analysis is that of the hazard function. In the absence of competing risks, the hazard function is defined as

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\text{prob}(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

The hazard function, which is a function of time, describes the instantaneous rate of occurrence of the event of interest in subjects who are still at risk of the event. In a setting in which the outcome was all-cause mortality, the hazard function at a given point in time would describe the instantaneous rate of death in subjects who were alive at that point in time. The Cox proportional hazards regression model relates the hazard function to a set of covariates. In the absence of competing events, the Cox proportional hazards regression model can be written as

$$\log(\lambda(t)) = \log(\lambda_0(t)) + X\beta$$

where $\lambda_0(t)$ denotes the baseline hazard function (ie, the hazard function for a subject whose covariates are all set equal to zero), $X$ denotes a set of explanatory
variables, and $\beta$ denotes the associated regression parameters. The model can also be written in multiplicative format: $\lambda(t) = \lambda_0(t) \exp(X\beta)$. The Cox model relates the covariates to the hazard function of the outcome of interest (and not directly to the survival times themselves). The covariates have a relative effect on the hazard function because of the use of the logarithmic transformation. The regression coefficients are interpreted as log-hazard ratios. The hazard ratio is equal to the exponential of the associated regression coefficient. The hazard ratio denotes the relative change in the hazard function associated with a 1-unit increase in the predictor variable. Although the regression coefficients from the Cox model describe the relative effect of the covariate on the hazard of the occurrence of the outcome, the following relationship also holds in the absence of competing risks' $S(t) = S_0(t)^{\exp(X\beta)}$, where $S(t)$ denotes the survival function for an individual whose set of covariates is equal to $X$, and $S_0(t)$ denotes the baseline survival function (i.e., the survival function for a subject whose covariates are all equal to zero). Thus, the relative effect of a given covariate on the hazard of the outcome is equal to the relative effect of that covariate on the logarithm of the survival function. Therefore, in the absence of competing risks, making inferences about the effect of a covariate on the hazard function permits one to make equivalent inferences about the effect of that covariate on prognosis or survival. This direct correspondence between the hazard function and incidence in the absence of competing risks has allowed authors to be imprecise in their language when interpreting the fitted Cox regression model.

Competing risks implies that a subject can experience one of a set of different events or outcomes. In this case, 2 different types of hazard functions are of interest: the cause-specific hazard function and the sub distribution hazard function. The former function is defined as

$$
\lambda_{cs}^k(t) = \lim_{\Delta t \to 0} \frac{\text{prob}(t \leq T < t + \Delta t, D = k \mid T \geq t)}{\Delta t}
$$

The cause-specific hazard function denotes the instantaneous rate of occurrence of the $k^{th}$ event in subjects who are currently event free (i.e., in subjects who have not yet experienced any of the different types of events). If one were considering 2 types of events, death attributable to cardiovascular causes and death attributable to non cardiovascular causes, then the cause-specific hazard of cardiovascular death denotes the instantaneous rate of cardiovascular death in subjects.
who have not yet experienced either event (ie, in subjects who are still alive). The sub
distribution hazard function, introduced by Fine and Gray, is defined as

$$\lambda_{k}^{sd}(t) = \lim_{\Delta t \to 0} \frac{\text{prob}(t \leq T + \Delta t, D = k \mid T > t \cup (T < t \cap k \neq k))}{\Delta t}$$

It denotes the instantaneous risk of failure from the $k^{th}$ event in subjects who have not yet
experienced an event of type $k$. Note that this risk set includes those who are currently event free
as well as those who have previously experienced a competing event. This differs from the risk
set for the cause specific hazard function, which only includes those who are currently event
free. Using the same example as above, the sub distribution hazard of cardiovascular death
denotes the instantaneous rate of cardiovascular death in subjects who are still alive (ie, who
have not yet experienced either event) or who have previously died of non cardiovascular causes.
There is a distinct cause-specific hazard function for each of the distinct types of events and a
distinct sub distribution hazard function for each of the distinct types of events.

**Inferential procedure**

Consider the usual competing risks setting where an event can be caused by $K$
competing risks, with $K \in \{1, 2, \ldots\}$ fixed. The random variables of interest are $(X, Y)$,
where $X \in \mathbb{R}$ is the time of the event of interest, and $Y \in \{1, \ldots, K\}$ is the
 corresponding cause. The goal is to estimate the cumulative incidence functions $F_0 = (F_{01},$
$\ldots, F_{0K})$, where $F_{0k}(t) = \text{pr}(X \leq t, Y = k)$ for $k = 1, \ldots, K$. The cumulative incidence
functions are non negative, monotone non decreasing, and satisfy

$$\sum_{k=1}^{K} F_{0k}(t) = \text{pr}(X \leq t) \leq 1.$$ 

The difficulty in estimating the cumulative incidence functions is that we cannot
observe $(X, Y)$ directly. Rather, we observe the current status of a subject at a single
random observation time $C \in \mathbb{R}$. Thus, at time $C$ we observe whether or not the event of
interest has occurred, and if and only if the event has occurred, we also observe the cause Y. We assume that C is independent of (X, Y). Let G denote the distribution of C, and let (C, Δ) denote the observed data, where Δ = (Δ₁, . . . , Δ_{K+1}) is an indicator vector for the status of the subject at time C: Δₖ = 1{X ≤ C, Y = k} (k = 1, . . . , K), Δ_{K+1} = 1{X > C}

The maximum likelihood estimator for F₀ based on n independent and identically distributed observations of (C,Δ), denoted by (Cᵢ, Δᵢ) (i =1, . . . , n), where Δᵢ = (Δᵢ¹, . . . , Δᵢ_{K+1}). For any K-tuple (x₁, . . . , xₖ) let x⁺ = \sum_{k=1}^{K} xₖ K and, unless otherwise defined, let 

X_{K+1}=1−x⁺. Moreover, define the set Fₖ = {F = (F₁, . . . , Fₖ) : F₁, . . . , Fₖ are cumulative incidence functions and F⁺(t) ≤ 1 for all t ∈ R}. A maximum likelihood estimator for F₀ is defined as any Fₙ =(Fₙ₁, . . . , Fₙₖ) ∈ Fₖ satisfying ln(Fₙ) = max F ∈ Fₖ ln(F), where

lₙ(F) is the log likelihood

\[ lₙ(F) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K+1} \Deltaₖ^{i} \log Fₖ(Cᵢ) \]

With the convention 0 log 0=0; refer to Jewell et al. (2003). The naive estimator Fₙ =(Fₙ₁, . . . , Fₙₖ) of Jewell et al. (2003), whose kᵗʰ component is defined as any Fₙₖ ∈ F₁ satisfying lₙₖ(Fₙₖ) = max F₁ ∈ F₁ lₙak(Fₙₖ), where

\[ lₙₖ(fₖ) = \frac{1}{n} \sum_{i=1}^{n} \left[ \Deltaₖ^{i} \log Fₖ(Cᵢ) + (1 - \Deltaₖ^{i}) \log (1 - Fₖ(Cᵢ)) \right] \]

Is the marginal log likelihood for the reduced current status data (Cᵢ, Δᵢₖ) (i =1, . . . , n), and F₁ is obtained from Fₖ by taking K =1. Since Fₙₖ only uses the kᵗʰ entry of the Δ-
vector, the naïve estimator splits the estimation problem into K well-known univariate current status problems.

Therefore, its computation and asymptotic theory follow straight forwardly from known results on current status data. But this simplification comes at a cost. For example, \( F_{n+} \) need not be bounded by unity, and the naive estimator has been empirically shown to be less efficient than the maximum likelihood estimator in the smooth model.

The R-package MLEcens provides an efficient and stable method to compute the maximum likelihood estimator. This algorithm first uses the Height Map Algorithm of Maathuis (2005) to compute the areas to which the maximum likelihood estimator can possibly assign probability mass, called maximal intersections. Next, it computes the amounts of mass that must be assigned to the maximal intersections. This involves solving a high-dimensional convex optimization problem, which is done using the support reduction algorithm of Groeneboom et al.(2008a). Jewell & Kalbfleisch (2004) describe an alternative algorithm for the computation of the maximum likelihood estimator, based on the pool adjacent violator’s algorithm of Ayer et al.(1955).

**Exact observation times with discrete support**

In this case, let \( G(\{s\}) \) denote the point mass of \( G \) at \( s \), and let

\[
S = \{ s \in R : G(\{s\}) > 0 \}
\]
denote the support of \( G \), where \( S \) is countable but possibly infinite.

Defining

\[
N_k(s) = \frac{1}{n} \sum_{i=1}^{n} \Delta_k' \mathbb{1}[C_i = s] \quad (k = 1, ..., K + 1), s \in S
\]

and

\[
N(s) = \sum_{k=1}^{K+1} N_k(s)
\]

the log likelihood (2) reduces to
\[ l_n(F) = \sum_{s \in S} \sum_{k=1}^{K+1} N_k(s) \log \{ F_k(s) \} \]

and the marginal loglikelihood (3) for the native estimator becomes

\[ \ln k(F_k) = \sum_{s \in S} [N_k(s) \log F_k(s)] + [N(s) - N_k(s)] \log [1 - F_k(s)] \]

The spaces \( F_k \) and \( F_1 \) can also be simplified, as the non negativity, monotonicity and boundedness constraints only need to hold at points \( s \in S \).

In many applications, only rounded versions of the observation times are recorded, yielding grouped observation times.

To derive the likelihood in the grouped model, continuous case we compute \( \Pr(D = d, \Delta = \delta) \) for \( d \in M \) and \( \delta \in \{ e_1, \ldots, e_{K+1} \} \), where \( e_k \) is the unit vector in \( \mathbb{R}^{K+1} \) with a 1 at the \( k^{th} \) entry. Conditioning on the exact observation time \( C \) yields.

\[
\Pr(D = d, \Delta = \delta) = \int \Pr(D = d, \Delta = \delta \mid C = c) dG(c) = \int_{c \in I(d)} \Pr(\Delta = \delta \mid C = c) dG(c)
\]

\[
= \prod_{k=1}^{K+1} \left\{ F_{0k}(c) dG(c) \right\}^{\delta \delta} = G(I(d)) \prod_{k=1}^{K+1} \left[ H_{0k}(I(d)) \right]^{\delta}
\]

where

\[
H_{0k}(I(d)) = [G(I(d))]^{-1} \int_{c \in I(d)} F_{0k}(c) dG(c) (k = 1, \ldots, K)
\]

and \( H_{0,K+1}(I(d)) = 1 - H_{0,+}I(d) \) are weighted averages of \( F_{01}, \ldots, F_{0,K+1} \) over \( I(d) \) with weights determined by \( G \).
The term \( G(I(d)) \) in the right hand side of (5) can be dropped from the likelihood, as it does not depend on \( F \). Hence, a maximum likelihood estimator for \( c \) is defined as any \( H_n \in H_k \) satisfying \( l_n^{\text{group}}(H_n) \) = \( \max_{H \in H_k} l_n^{\text{group}}(H) \), where

\[
l_n^{\text{group}}(H) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K+1} \Delta_i^k \log[H_k(I(D_i))]\]

Expression has the same form as, but with \( F_k(C_i) \) replaced by the weighted average. As in the discrete model, can be simplified further:

\[
l_n^{\text{group}}(H) = \sum_{i=1}^{n} \sum_{k=1}^{K+1} M_k(I) \log[H_k(I)]\]

Where

\[
M_k(I) = \frac{1}{n} \sum_{i=1}^{n} \Delta_i^k 1[D_i = m(I)](k = 1,..,K+1), I \in I
\]

Since the log likelihood (7) has the same form as (4), and also the constraints on the maximization

Problems for the discrete and grouped models are equivalent; the maximum likelihood estimator in the grouped model can be computed with existing software. Moreover, its asymptotic theory follows straightforwardly from the theory for the discrete model. The important difference between the two models is, however, that the resulting estimates must be interpreted differently. In the discrete model, one estimates the cumulative incidence functions at point’s \( s \in S \). In the grouped model, the cumulative incidence
functions are unidentifiable in general, and one estimates the weighted averages of the cumulative incidence functions over intervals \( I \in I \).

The naive estimator \( \hat{H}_n \) in the grouped model can be derived analogously.

Defining \( M(I) = \sum_{k=1}^{K+1} \sum_{i=1}^{n} [M_k(I) \log(H_k(I) - M_k(I))] \),

the marginal log likelihood for the \( k \)th component is

\[
\log\{\hat{H}_{nk} \in H_1 \text{ is defined by } l_{nk}^{\text{group}}(\hat{H}_{nk}) = \max_{H_k \in H_1} l_{nk}^{\text{group}}(H_k) \}
\]

Confidence intervals in the discrete and grouped models

In the discrete and grouped models, the large-sample behavior of the maximum likelihood estimator and the naive estimator at regular points or intervals is standard, and hence confidence intervals can be constructed by any standard method, for example using the asymptotic normal distribution or the bootstrap. For instance, let \( s \in S \) be a regular point in the discrete model. Then an asymptotic \((1 - \alpha)100\%\) confidence interval for

\[
F_{0k}(s)
\]

is

\[
\hat{F}_{nk}(s) \pm n^{-1/2} \alpha / 2 [V_s(s)]^{1/2}
\]

Where \( Z_{1-\alpha/2} \) is the \((1 - \alpha/2)\) quantile of the standard normal distribution.

Similarly, considering a regular interval \( I \in I \) in the grouped model, an asymptotic
\( (1 - \alpha) 100\% \) confidence interval \( H_{nk}(I) \)

\[
\hat{H}_{nk}(I) \pm n^{-1/2} z_{1 - \alpha / 2} [\hat{U}_n(I)_{k,k}]^{1/2}
\]

**Example**

**(i) MENOPAUSE DATA**

We consider data on 2423 women in the age range 25–59 years from Cycle I of the Health Examination Survey of the National Center for Health Statistics (MacMahon & Worcester, 1966). Among other things, these women were asked to report: their current age; whether they were pre- or postmenopausal; and if they were postmenopausal, the age and cause of menopause, where the cause could be natural or operative. Since MacMahon & Worcester (1966) found marked terminal digit clustering in the reported ages of menopause, Krailo & Pike (1983) excluded these from the analysis. The remaining information can be viewed as current status data with competing risks. Nonparametric estimates of the cumulative incidences of the two types of menopause were computed by Jewell et al. (2003) and Jewell & Kalbfleisch (2004) under the assumption that the recorded ages of the women at the time of the interview were exact. However, this was not the case. Instead, the ages were grouped into the intervals 25–30, 30–35, 35–36, 36–37 . . . , 58–59 and recorded as the midpoints of these intervals, yielding 26 age groups with a minimum of 45 and an average of 93 observations per age group.
(ii) HIV data

The Bangkok Metropolitan Administration injecting drug user’s cohort study (Kitayaporn et al., 1998; Vanichseni et al., 2001) was established in 1995 to better understand HIV transmission and to assess the feasibility of conducting a phase III HIV vaccine efficacy.

Fig.1. Simulation: Coverage (a)–(d) and average width (e)–(h) of the four 95% confidence intervals for $F_{01(t_0)}$ as a function of $t_0$, corresponding to grids gap 10 { (a) and (e) }, gap 2 { (b) and (f) }, gap 0.5 { (c) and (g) }, and gap 0.1 { (d) and (h) }. The confidence intervals were based on the normal distribution (◦) and the bootstrap (Δ), using the
maximum likelihood estimator (solid line) and the naive estimator (dashed line). The bootstrap confidence intervals are based on 750 bootstrap samples.

Fig. 2. Menopause data. The maximum likelihood estimator $\hat{H}_n (\circ)$ and the naive estimator $\hat{H}_n (\times)$ for the weighted averages of the cumulative incidence of (a) operative and (b) natural menopause over the age groups. The estimators are plotted at the midpoints of the age groups which are indicated by the dotted vertical lines. The two solid vertical line segments in each age group are 95% asymptotic confidence intervals based on the maximum likelihood estimator: the left line segment is based on the normal approximation (8) and the right line segment is a symmetric bootstrap confidence interval based on 1000 bootstrap samples. trial in an injecting drug user’s population in Bangkok. We consider data on 1366 injecting drug users in this study who were screened from May to December 1996 and who were under 35 years of age. Among this group, 393 were
HIV positive, with 114 infected with subtype B, 238 infected with subtype E, 5 infected by another or mixed subtype and 36 infected with missing subtype.

The subjects with other, mixed or missing subtypes were grouped in a remainder category. All ages were recorded in days, leading to a small number of ties: among the 1366 subjects, there were 1212 distinct ages, and the mean number of observations per distinct age was 1.13. In light of this, we analyze these data using the smooth model.

Figure 3 shows the maximum likelihood estimator and the naive estimator for the subtype-specific cumulative incidence of HIV, together with 95% likelihood ratio confidence intervals based on the naive estimator.

Competing risks occur frequently in the analysis of survival data. A competing risk is an event whose occurrence precludes the occurrence of the primary event of interest.

The Kaplan-Meier method for estimating survival functions and the Cox proportional hazards model for estimating the effects of covariates on the hazard of the occurrence of the event are commonly used statistical methods for the analysis of survival data.

**Analysis of Survival Data in the Presence of Competing Risks:**

**Estimating Crude Incidence**

We assume that there is a well-defined baseline time in the cohort and that \( T \) denotes the time from baseline time until the occurrence of the event of interest. In the absence of competing risks, the survival function, \( S(t) \), describes the distribution of event
times: \( S(t) = \Pr(T > t) \). One minus the survival function (ie, the complement of the survival function), \( F(t) = 1 - S(t) = \Pr(T \leq t) \) describes the incidence of the event over the duration of follow-up. Two key properties of the survival function are that \( S(0) = 1 \) (ie, at the beginning of the study, the event has not yet occurred for any subjects) and 
\[
\lim_{t \to \infty} S(t) = 0 \quad \text{(ie, eventually the event of interest occurs for all subjects). In practice, the latter assumption may not be required, because the probability of the event over a restricted follow-up period may be <1.}
\]

Estimating the incidence of an event as a function of follow-up time provides important information on the absolute risk of an event. In the absence of competing risks, the Kaplan-Meier estimate of the survival function is frequently used for estimating the survival function. One minus the Kaplan-Meier estimate of the survival function provides an estimate of the cumulative incidence of events over time.

The Cumulative Incidence Function (CIF), as distinct from \( 1 - S(t) \), allows for estimation of the incidence of the occurrence of an event while taking competing risk into account. This allows one to estimate incidence in a population where all competing events must be accounted for in clinical decision making. The cumulative incidence function for the \( k^{th} \) cause is defined as: \( \text{CIF}_k(t) = \Pr(T \leq t, D = k) \), where \( D \) is a variable denoting the type of event that occurred. A key point is that, in the competing risks setting, only 1 event type can occur, such that the occurrence of 1 event precludes the subsequent occurrence of other event types. The function \( \text{CIF}_k(t) \) denotes the probability of experiencing the \( k^{th} \) event before time \( t \) and before the occurrence of a different type of
event. The CIF has the desirable property that the sum of the CIF estimates of the incidence of each of the individual outcomes will equal the CIF estimates of the incidence of the composite outcome consisting of all of the competing events.

**Hazard Function Regression**

A key concept in survival analysis is that of the hazard function. In the absence of competing risks, the hazard function is defined as

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\Pr(\text{event} \leq t + \Delta t | T \geq t)}{\Delta t}$$

The hazard function, which is a function of time, describes the instantaneous rate of occurrence of the event of interest in subjects who are still at risk of the event. In a setting in which the outcome was all-cause mortality, the hazard function at a given point in time would describe the instantaneous rate of death in subjects who were alive at that point in time. The Cox proportional hazards regression model relates the hazard function to a set of covariates. In the absence of competing events, the Cox proportional hazards regression model can be written as

$$\log(\lambda(t)) = \log(\lambda_0(t)) + X\beta$$

where \(\lambda_0(t)\) denotes the baseline hazard function (i.e., the hazard function for a subject whose covariates are all set equal to zero), \(X\) denotes a set of explanatory variables, and \(\beta\) denotes the associated regression parameters. The model can also be written in multiplicative format:

$$\lambda(t) = \lambda_0(t) \exp(X\beta)$$

The Cox model relates the covariates to the hazard function of the outcome of interest (and not directly to the survival times themselves). The covariates have a relative effect on the hazard function because of the use of the logarithmic transformation. The regression coefficients are interpreted as log-hazard ratios. The
hazard ratio is equal to the exponential of the associated regression coefficient. The hazard ratio denotes the relative change in the hazard function associated with a 1-unit increase in the predictor variable. Although the regression coefficients from the Cox model describe the relative effect of the covariates on the hazard of the occurrence of the outcome, the following relationship also holds in the absence of competing risks:

\[ S(t) = S_0(t)^{\exp(x)} \]

where \( S(t) \) denotes the survival function for an individual whose set of covariates is equal to \( X \), and \( S_0(t) \) denotes the baseline survival function (i.e., the survival function for a subject whose covariates are all equal to zero). Thus, the relative effect of a given covariate on the hazard of the outcome is equal to the relative effect of that covariate on the logarithm of the survival function. Therefore, in the absence of competing risks, making inferences about the effect of a covariate on the hazard function permits one to make equivalent inferences about the effect of that covariate on prognosis or survival. This direct correspondence between the hazard function and incidence in the absence of competing risks has allowed authors to be imprecise in their language when interpreting the fitted Cox regression model.

Competing risks implies that a subject can experience one of a set of different events or outcomes. In this case, 2 different types of hazard functions are of interest: the cause-specific hazard function and the sub distribution hazard function. The former function is defined as

\[ \hat{\lambda}_k^X(t) = \lim_{\Delta t \to 0} \frac{\Pr\{b(t \leq T + \Delta t, D = k \mid T > t)\}}{\Delta t} \]
The cause-specific hazard function denotes the instantaneous rate of occurrence of the $k$th event in subjects who are currently event free (i.e., in subjects who have not yet experienced any of the different types of events). If one were considering 2 types of events, death attributable to cardiovascular causes and death attributable to noncardiovascular causes, then the cause-specific hazard of cardiovascular death denotes the instantaneous rate of cardiovascular death in subjects who have not yet experienced either event (i.e., in subjects who are still alive). The sub distribution hazard function, introduced by Fine and Gray, is defined as

$$\lambda_k^c(t) = \lim_{\Delta t \to 0} \frac{Prob(t < T \leq +\Delta t, D = k \mid T > t \cup (T < t \cap K \neq k))}{\Delta t}$$

It denotes the instantaneous risk of failure from the $k$th event in subjects who have not yet experienced an event of type $k$. Note that this risk set includes those who are currently event free as well as those who have previously experienced a competing event. This differs from the risk set for the cause specific hazard function, which only includes those who are currently event free. Using the same example as above, the Downloaded from sub distribution hazard of cardiovascular death denotes the instantaneous rate of cardiovascular death in subjects who are still alive (i.e., who have not yet experienced either event) or who have previously died of noncardiovascular causes. There is a distinct cause-specific hazard function for each of the distinct types of events and a distinct sub distribution hazard function for each of the distinct types of events.
Statistical Software for Competing Risks Analyses

CIFs can be estimated in R using the cuminc function in the cmprsk package; in SAS, one can use the %CIF macro; Stata permits estimation of the CIF using the st curve function. Note that in SAS/STAT 13.1, %CIF is an auto call macro, and thus does not need to be loaded manually by the analyst. Cause-specific hazard models can be fit in any statistical software package that permits estimation of the conventional Cox proportional hazards model. One simply treats those subjects who experience a competing event as being censored at the time of the occurrence of the competing event. In R, one can use the coxph function in the survival package, in SAS, one can use PROC PHREG, and in Stata, one can use the stcox function.

Sub distribution hazard models can be fit in R by using the crr function in the cmprsk package. In SAS, PROC PHREG permits estimation of sub distribution hazard models through the use of the 'event code' option in the model statement (in SAS/STAT version 13.1). In Stata, the stcrreg function permits estimation of sub distribution hazard regression models. In the case study below, we used the R (version 3.1.2) statistical programming language and the cmprsk package (version 2.2–6) for all of the statistical analyses. R code for estimating the CIFs, the sub distribution hazard models and the cause-specific hazard models is described in Appendix A in the online-only Data Supplement. SAS code for fitting these functions and models is described in Appendix B in the online-only Data Supplement.
Illustration

The Enhanced Feedback for Effective Cardiac Treatment (EFFECT) Study was designed to assess the effect of public reporting of hospital performance on the quality of care provided to patients with cardiovascular disease in Ontario, Canada. We obtained detailed clinical data by retrospective chart review on patients hospitalized with heart failure (HF) between April 1, 1999 and March 31, 2001 (phase 1) and between April 1, 2004 and March 31, 2005 (phase 2) at 103 hospitals in Ontario, Canada. Trained cardiovascular nurse abstractors collected data on patient demographics, vital signs and physical examination at presentation, medical history, and results of laboratory tests. These data sets were linked by using unique, encoded identifiers and they were analyzed at the Institute for Clinical Evaluative Sciences. We considered 11 baseline covariates, which make up the EFFECT-HF mortality prediction model: age, systolic blood pressure on admission, respiratory rate on admission, low sodium serum concentration (<136 mEq/L), low serum hemoglobin (<10.0 g/dL), serum urea nitrogen, presence of cerebrovascular disease, presence of dementia, chronic obstructive pulmonary disease, hepatic cirrhosis, and cancer. The initial sample consisted of 18 284 patients hospitalized with HF. We excluded 255 subjects who were on dialysis for end-stage renal failure, because the EFFECT-HF mortality prediction model was not intended for use in these subjects. We excluded an additional 1792 subjects with missing data on continuous covariates in the EFFECT-HF mortality prediction model. This left 16 237 patients for analysis (8521 patients in phase 1 and 7716 patients in phase 2). Subjects were linked by using an encoded version of the patient’s Ontario health insurance number to the Vital...
Statistics database maintained by the Ontario Office of the Registrar General. This database contains information on date of death and cause of death for residents of Ontario. Each subject was followed for 5 years from the date of hospital admission for the occurrence of death. For those subjects who died within 5 years of discharge, the cause of death was noted in the Vital Statistics database. We categorized cause of death as cardiovascular versus noncardio vascular: 10 215 (63%) patients died during the 5 years of follow-up. Of these, 5970 (58%) died of cardiovascular causes, and 4245 (42%) died of noncardio vascular causes.

**Hazard Models for Cardiovascular Death**

We fit cause-specific and sub distribution hazard models for both cardiovascular death and noncardiovascular death. For each of the 2 causes of death, we regressed the hazard of death on the 11 covariates described above. The estimated hazard ratios, along with their associated confidence intervals, are reported in Table 2.

**Table -1**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Effect - HF sample (N = 16237)</th>
<th>Cardiac Death (N=5970)</th>
<th>Non cardiac Death (N=4245)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>78(70-84)</td>
<td>81(74-87)</td>
<td>79(73-85)</td>
</tr>
<tr>
<td>Respiratory rate at admission, breaths / min</td>
<td>24(20-28)</td>
<td>24(20-28)</td>
<td>24(20-28))</td>
</tr>
<tr>
<td>Systolic blood pressure at admission, mm Hg</td>
<td>145 (124-169)</td>
<td>140(120-162)</td>
<td>144(124-168)</td>
</tr>
<tr>
<td>Serum urea nitrogen, mg/dL-</td>
<td>24(17-34))</td>
<td>27(19-39)</td>
<td>25(18-38)</td>
</tr>
<tr>
<td>Cancer</td>
<td>1889(12)</td>
<td>631(11)</td>
<td>764(18)</td>
</tr>
<tr>
<td>Cirrhosis</td>
<td>120(1)</td>
<td>44(1)</td>
<td>46(1)</td>
</tr>
<tr>
<td>Cerebro vascular disease</td>
<td>2827(17)</td>
<td>1267(21)</td>
<td>784(18)</td>
</tr>
<tr>
<td>Dementia</td>
<td>1457(9)</td>
<td>743(12)</td>
<td>510(12)</td>
</tr>
<tr>
<td>Chronic obstructie pulmonary disease</td>
<td>4066(25)</td>
<td>1499(25)</td>
<td>1316(31)</td>
</tr>
</tbody>
</table>
Cells reporting patient characteristics contain either n(%) for dichotomous variables or median (25th percentile-75th percentile (Q1-Q3) for continuous variables. Effect - HF indicates Enhanced Feedback for Effective Cardiac Treatment Heart Failure Study.

**Figure 1.** Cumulative incidence functions. CIF incidence function; and KM, Kaplan-Meier.

A 10-year increase in age increased the relative incidence of cardiac death by 42%, whereas it increased the relative incidence of noncardiac death by 14%. Similarly, a 10-year increase in age increased the cause-specific hazard of cardiac death by 52%, whereas it increased the cause-specific hazard of noncardiac death by 31%. Thus, age had a more pronounced effect on both the incidence and cause-specific hazard of cardiac mortality than on noncardiac mortality. Furthermore, age had a more pronounced effect...
on the cause-specific hazard of a given outcome than it did on the incidence of the same outcome. The presence of cancer had a small and non significant effect on the cause-specific hazard of cardiac death, whereas

**Figure 2.** Cumulative incidence functions and Kaplan-Meier estimates. CIF indicates cumulative incidence function; and KM.
### Table - 2

Baseline characteristics of Patients in the Effect HF sample

<table>
<thead>
<tr>
<th>Variable</th>
<th>Sub distribution Hazard Model</th>
<th>Cause-Specific Hazard Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cardiac Death</td>
<td>Non Cardiac Death</td>
</tr>
<tr>
<td>Age (PER 10-y-increase in age)</td>
<td>1.42(1.38-1.46)</td>
<td>1.14(1.10-1.17)</td>
</tr>
<tr>
<td>Respiratory rate (per 10 breaths /min increase)</td>
<td>1.12(1.08-1.17)</td>
<td>1.07(1.02-1.12)</td>
</tr>
<tr>
<td>Systolic blood pressure (per 10mm Hg increase)</td>
<td>0.93(0.92-0.94)</td>
<td>1.00(0.99-1.01)</td>
</tr>
<tr>
<td>Urea nitrogen</td>
<td>1.02(1.02-1.02)</td>
<td>1.01(1.01-1.01)</td>
</tr>
<tr>
<td>Cancer</td>
<td>0.82(0.75-0.89)</td>
<td>1.85(1.71-2.01)</td>
</tr>
<tr>
<td>Cirrhosis</td>
<td>1.12(0.82-1.54)</td>
<td>1.49(1.10-2.02)</td>
</tr>
<tr>
<td>Cerebrovascular disease</td>
<td>1.26(1.18-1.35)</td>
<td>1.01(0.93-1.09)</td>
</tr>
<tr>
<td>Dementia</td>
<td>1.22(1.11-1.33)</td>
<td>1.35(1.22-1.50)</td>
</tr>
<tr>
<td>COPD</td>
<td>1.03(0.97-1.1)</td>
<td>1.42(1.33-1.52)</td>
</tr>
<tr>
<td>Low Hemoglobin</td>
<td>0.87(0.80-0.94)</td>
<td>1.49(1.37-1.62)</td>
</tr>
<tr>
<td>Low sodium concentration</td>
<td>1.14(1.07-1.22)</td>
<td>1.13(1.05-1.22)</td>
</tr>
</tbody>
</table>

Each cell contains the hazard ratio and associated 95% confidence interval for the given covariate and the given hazard model. COPD indicates chronic obstructive pulmonary disease.

It increased the cause-specific hazard of noncardiac death by 85%. This illustrates how the apparent reduction in the absolute risk of cardiac death from cancer may be explained via the effect of cancer on noncardiac death.

In examining the estimated hazard ratios for the different outcomes and different types of hazard models, one notes that some variables have a qualitatively similar effect.
on the incidence of cardiac death as on the incidence of noncardiac death. However, other variables have a qualitatively different effect on the incidence of cardiac death in comparison with the effect on the incidence of noncardiac death. Furthermore, some variables have a qualitatively similar effect on the incidence of a given type of death as on the cause-specific hazard for the same type of death. As described elsewhere, the interpretation of these results requires careful consideration of the effects of variables on competing causes of death. As an example, a strong and opposing effect of a variable on the cause-specific hazard of a competing event may lead to an indirect effect on the cumulative incidence of the event of interest. To be concrete, a strong prognostic factor for the cause-specific hazard for cardiovascular death might lead to an apparent decrease in the cumulative incidence for noncardiovascular death when such factor has no effect on the cause specific hazard for noncardiovascular death. This indirect effect of the prognostic factor for cardiovascular death occurs because noncardiovascular death cannot occur in those who die of cardiovascular causes and hence have a decreased risk for that event.

Conclusion

The analysis of survival data plays a key role in cardiovascular research. Competing risks are prevalent in much of cardiovascular research. Failure to account correctly for competing events can result in adverse consequences, including overestimation of the probability of the occurrence of the event and miss-estimation of the magnitude of relative effects of covariates on the incidence of the outcome. Koller et al15 found that competing risks were present in a large majority of studies published in a
sample of high-impact journals. This suggests that it is crucial that investigators be aware of appropriate methods to account for competing risks when analyzing survival data. We have provided a brief, nontechnical, introduction to statistical methods to account for the presence of competing risks. When estimating crude incidence of the outcome of interest, it is inappropriate to use the complement of the Kaplan-Meier survival function, because this will result in an overestimate of the incidence of the outcome of interest when competing risks are present. Instead, authors and analysts are encouraged to use the CIF. Competing risks are prevalent in cardiovascular research. We encourage analysts to take full advantage of the range of statistical methods for the analysis of survival data that have been developed in the statistical literature. Investigators need to be cognizant of the presence of competing risks and their potential effect on statistical analyses. Researchers should select the appropriate method to address the study objectives and ensure that the analysis results are interpreted correctly.

REFERENCES


