Special Section

A Review of Methods to Estimate Cause-Specific Mortality in Presence of Competing Risks

DENNIS M. HEISEY, *United States Geological Survey National Wildlife Health Center, Madison, WI 53711, USA*

BRENT R. PATTERSON,[1] *Ontario Ministry of Natural Resources, Wildlife Research and Development Section, Peterborough, ON K9J 8M5, Canada*

Abstract

Estimating cause-specific mortality is often of central importance for understanding the dynamics of wildlife populations. Despite such importance, methodology for estimating and analyzing cause-specific mortality has received little attention in wildlife ecology during the past 20 years. The issue of analyzing cause-specific, mutually exclusive events in time is not unique to wildlife. In fact, this general problem has received substantial attention in human biomedical applications within the context of biostatistical survival analysis. Here, we consider cause-specific mortality from a modern biostatistical perspective. This requires carefully defining what we mean by cause-specific mortality and then providing an appropriate hazard-based representation as a competing risks problem. This leads to the general solution of cause-specific mortality as the cumulative incidence function (CIF). We describe the appropriate generalization of the fully nonparametric staggered-entry Kaplan–Meier survival estimator to cause-specific mortality via the nonparametric CIF estimator (NPCIFE), which in many situations offers an attractive alternative to the Heisey–Fuller estimator. An advantage of the NPCIFE is that it lends itself readily to risk factors analysis with standard software for Cox proportional hazards model. The competing risks–based approach also clarifies issues regarding another intuitive but erroneous "cause-specific mortality" estimator based on the Kaplan–Meier survival estimator and commonly seen in the life sciences literature. (JOURNAL OF WILDLIFE MANAGEMENT 70(6):1544–1555; 2006)

Key words

Canis latrans, *Cox proportional hazards, coyote, cumulative incidence function, Kaplan–Meier, MICROMORT, product limit estimator, survival analysis.*

During the past 20 years, there has been much elaboration of methods available for estimation of survival rates and distributions for wildlife populations (Winterstein et al. 2001, DelGiudice et al. 2002, Williams et al. 2002). This has arisen largely by adaptation of methods used in human biomedical applications, mark–recapture estimation of animal densities, or industrial failure-time estimation. Relative to recent advances in survival analysis, researchers have directed less attention to the estimation and comparison of cause-specific mortality (CSM) rates, despite the importance of this information for improvement in our understanding of the dynamics of wildlife populations. The issue of analyzing cause-specific, mutually exclusive events has received substantial attention in human biomedical applications within the context of biostatistical survival analysis. Here, we attempt to update analyses of cause-specific mortality in wildlife research from a modern biostatistical perspective.

Mortality and survival rates are actually probabilities, and probabilities are associated with random variables. For mortality and survival probabilities, the random variable is survival time (T). The mortality probability at time t, $M(t)$, is the probability that the survival time T will be less than some specified time t, which is usually represented as $M(t) = \Pr(T < t)$. Survival probability is the complement of mortality probability, that is $S(t) = 1 - M(t) = \Pr(T \geq t)$. Cause-specific mortality introduces a second random variable, K, which is the type (cause) of death. Thus, cause-specific mortality involves the bivariate pair of random variables T and K. Cause-specific mortality is the joint probability that death occurs before t, $T < t$, and that it is of type $K = k$. We represent this symbolically as $M_k(t) = \Pr(T < t, K = k)$. Because the causes, k, are mutually exclusive events (assume there are Q different causes), an important and natural property of this definition of cause-specific mortality is that the cause-specific mortality probabilities sum to the total mortality probability:

$$M(t) = \sum_{k=1}^{Q} M_k(t)$$

We will refer to this property as "conservation of mortality." All reasonable estimators of cause-specific mortality should obey this property, and one should be skeptical of estimators that exhibit substantial violations.

If researchers mark a sample of animals simultaneously at the start of a period of interest, calculation of $S(t)$ and $M_k(t)$ as simple proportions (e.g., binomial estimator, see Murray 2006) is appropriate provided no subjects are lost to follow-up (i.e., censored), except due to a known death. However, if researchers progressively enter animals into the at-risk group as a study progresses (i.e., staggered entry sensu Pollock et al. 1989, more commonly known in biostatistics as left-truncation) then 2 serious biases may arise (Heisey and Fuller 1985):

1. When researchers periodically add animals to the risk set as a study progresses, unmarked animals that die early during the sampling interval are not available for addition

[1] *E-mail: brent.patterson@mnr.gov.on.ca*

to the risk set later in the sampling interval. If not properly accounted for, pooling of animals added part way through the study with animals included in the risk set since the beginning of the study biases upwards estimated survival rates for the entire study period, and hence underestimates the magnitude of specific mortality agents.

2. If survival probability varies seasonally, then seasons having the largest sample sizes will be most influential in determining overall survival rates. This occurs because using a simple binomial estimator of survivor, researchers calculate annual mortality by dividing the number of deaths observed during the entire year by the number of marked animals "at risk" during the year, regardless of the season or period of the year during which they were actually part of the risk set. For sake of example, assume there are no seasonal differences in survival probability but that sample sizes are highest in winter following capture of new study animals. Overall, most mortalities documented during the study will occur in winter (simply because of higher sample sizes in winter) and because data from animals sampled throughout the year are pooled and considered equal with a simple proportion estimator, winter will be most influential in determining annual survival rates.

As noted, these simple proportion estimators also assume that no marked animals drop out during the period of interest (i.e., no right censoring). Prior to Heisey and Fuller (1985), researchers often expressed cause-specific mortality rates as simple proportions or percentages and thus were subject to the biases discussed above.

Heisey and Fuller (HF; 1985) developed an estimator of $M_k(t)$ based on Trent and Rongstad's (1974) life table-like approach to estimating $M(t)$. Trent and Rongstad's approach is equivalent to assuming a piecewise constant hazard function, which is a popular approach in mainstream biostatistical survival analysis referred to as semiparametric (Laird and Oliver 1980) or loosely parametric (Cai and Betensky 2003). The primary criticism of this approach is the arbitrariness of interval construction. For wildlife applications, Pollock et al. (1989) popularized a fully nonparametric estimator of $M(t)$ (i.e., makes no assumptions about the shape of the underlying distribution), known in biostatistical survival analysis as the generalized Kaplan–Meier (1958; GKME) or Lynden-Bell (1971) estimator. Although it has seen little application in wildlife ecology, the Breslow nonparametric survival estimator is often an attractive alternative to GKME (Therneau and Grambsch 2000). Nonetheless, we focus here primarily on the GKME. In certain situations, fully parametric models of $M(t)$, such as the Weibull or Gompertz, correctly formulated to account for left-truncation (e.g., Cox and Oakes 1984), may also be useful (see Murray 2006). For any valid estimator of $M(t)$, it is worthwhile to consider how to extend it to become a valid estimator of cause-specific mortality $M_k(t)$. Because of their assumption-free nature, generalizations of fully nonparameteric estimators of $M(t)$ such as the GKME and Breslow estimator are especially interesting.

We focus here on the fully nonparametric estimator of $M_k(t)$ that arises from the GKME and has an appealingly simple form that obeys the conservation of mortality property. We refer to this estimator specifically as the nonparametric cumulative incidence function estimator (NPCIFE) but we recognize that the Breslow or other survival estimators could be the basis for equally valid nonparametric CIF estimators. As we discuss in more detail below, the NPCIFE is not equivalent to the "common-sense" approach of estimating cause-specific mortality by treating all deaths except those due to cause k as censored events. This latter approach is particularly common in the human epidemiological field, referred to as the "complement of the Kaplan–Meier method" (1-KM) by Pepe and Mori (1993) and Gooley et al. (1999), and later coined "the KM ignore method (KM-I)" by Tai et al. (2001). The 1-KM approach has also appeared in the wildlife literature, for example, Kurzejeski et al. (1987) erroneously used this estimator to determine "cause-specific mortality" in wild turkeys. The 1-KM approach estimates the probability of an event of type k by time t, in the *absence* of competing risks and thus tends to overestimate the magnitude of specific mortality causes when multiple sources of mortality (competing risks) are present (Gaynor et al. 1993, Gooley et al. 1999, Tai et al. 2001). We show why the NPCIFE, and not 1-KM, is the appropriate generalization of the GKME for estimating $M_k(t)$. For overall mortality or survival, Cox proportional hazards model (Cox 1972) has become popular for examining the influence of various covariates or risk factors. We will demonstrate how to easily obtain the NPCIFE as an extension of Cox-type analyses for cause-specific mortality.

In considering the NPCIFE and HF approaches, one encounters an apparent paradox. Heisey and Fuller (1985), arguing on the basis of parsimony, suggest that researchers should choose intervals judiciously and combine apparently similar intervals. Researchers sometimes criticize the HF approach because arbitrary interval selection can lead to potentially biased survival or mortality rate estimates (e.g., Tsai et al. 1999). To avoid this in HF, the analyst could allow the maximum possible number of intervals (e.g., every day is a new interval), at which point the HF estimator becomes the NPCIFE. However, doing this contradicts the importance of parsimony argued for by Heisey and Fuller (1985).

This apparent paradox is resolved by carefully considering what constitutes a good estimator. A good estimator is one that in any particular application is likely to be close to the truth. A common misconception in wildlife statistics and elsewhere is that bias alone is a useful measure of likely closeness to the truth. In the text *Statistical Theory*, Lindgren (1976:255) reflects modern statistical thinking on this issue when he states:

> "...the everyday connotation of the words 'bias' and 'unbiased' have led to a common belief—even convic-

tion—that an estimator should be unbiased to be any good. There is clearly no basis for this belief as a general rule."

In modern biostatistics, the standard measure for likely closeness to the truth is the mean squared error (MSE), which reflects the compromise between variance and bias. We briefly consider how MSE considerations, and neither bias nor variance alone, should determine the choice of cause-specific mortality estimator. We also discuss future directions of cause-specific mortality, specifically procedures that avoid the variance and identifiability issues of strictly nonparametric approaches applied to left-truncated data (Tsai et al. 1987) and the arbitrariness of piecewise models.

Cumulative Incidence Functions

Death can occur at essentially any instant in time, so it is appropriate to treat the time at death T as a continuous random variable. This forms the foundation of modern biostatistical survival analysis. The use of discrete-time approximations are popular in wildlife survival analyses, but a continuous time approach to cause-specific mortality is unavoidable to present a sufficient characterization of the issues that need to be addressed.

We must assume that researchers monitor the subject animals frequently enough that they can estimate time of death T with negligible error. If this assumption cannot be met, researchers must employ methods for interval-censored or grouped data (e.g., Heisey et al. 2006). The treatment of cause-specific mortality in this context is quite complex and beyond the scope of this paper.

A key concept is the notion of a continuous time hazard function, or force-of-mortality. The hazard function, $h(t)$, is arguably the most important quantity in modern survival analysis and gives the instantaneous probability per unit time for death (failure) to occur. Suppose we are interested in the probability of death due to cause k over some very short time interval, Δt. Symbolically, this is written as $\Pr(t < T < t + \Delta t, K = k | T > t) = m_k(t + \Delta t | t)$, which means, conditional on being alive at time t ("$|T > t$"), death occurs in the time interval t to $t + \Delta t$ ("$t < T < t + \Delta t$"), and that the death is due to cause k. To model this, we define the cause-specific hazard rate function $h_k(t)$ which behaves such that

$$m_k(t + \Delta t | t) \approx h_k(t)\Delta t \tag{1}$$

That is, over very short intervals of time, researchers can model mortality as acting in a very simple linear fashion. Because the interval is assumed to be so short that the probability of two events occurring is essentially zero, we can represent the probability of dying due to any cause as the sum of the hazards $m(t + \Delta t | t) \approx [h_1(t) + \ldots + h_Q(t)]\Delta t$. The function $h_k(t)$ is the cause-specific hazard function, formally defined as the limit of Eq. 1 as $\Delta t \to 0$. By working with processes that are occurring at an instant in time, one can work with simple additive or multiplicative relationships instead of the complicated exponential expressions involved in finite survival probabilities. This is why hazard functions form the basis for most survival analysis modeling.

Suppose we divide time into fine intervals with boundaries at $t_0 (= 0) < t_1 < t_2 < \ldots < t_M$. The multiplicative nature of survival across time intervals is familiar from life table applications (e.g., Banfield 1955, Caughley 1966). The interval mortalities $m(t_{i+1} | t_i)$ can be easily combined using the multiplicative relationship to obtain the overall survival:

$$S(t) = \prod_{i=0}^{M} [1 - m(t_{i+1} | t_i)] \approx \prod_{i=0}^{M} [1 - h(t_i)(t_{i+1} - t_i)] \tag{2}$$

We can divide these intervals in increasingly more and finer intervals, eventually reaching the limit, which we represent as

$$S(t) = \prod_{i \to \infty} [1 - m(t_{i+1} | t_i)] \approx \prod_{i \to \infty} [1 - h(t_i)(t_{i+1} - t_i)] \tag{3}$$

a Volterra product-integral (Gill and Johansen 1990). Just as the usual Riemann integral is the limit of a series of sums, the Volterra product-integral is the limit of a series of products (Anderson et al. 1993). The solution to this integral shows the relationship of the hazard function $h(t)$ to the survival function $S(t)$:

$$S(t) = \exp\left[-\int_0^t h(u)du\right] \tag{4}$$

Because of the additivity of cause-specific hazards, we can express this in terms of the cause-specific hazards as

$$S(t) = \exp\left[-\int_0^t \{h_1(u) + \ldots + h_Q(u)\} du\right] \tag{5}$$

If we define

$$S_k(t) = \exp\left[-\int_0^t \{h_k(u)\} du\right] \tag{6}$$

we can then write overall survival as

$$S(t) = \prod_{k=1}^{Q} S_k(t) \tag{7}$$

$S_k(t)$ will become important later in the discussion. It is important to understand that $S_k(t)$ does not have an interpretation as a observable, realizable quantity but is simply a useful way to algebraically decompose $S(t)$. Recall that $h_k(t)$ is the mortality hazard due to cause k, with all $Q - 1$ other causes of mortality present and operating. We can view $S_k(t)$ as the survival expected if we removed all sources of mortality except k, under the assumption that the removal of all other sources of mortality would not influence the hazard $h_k(t)$. This assumption is both unverifiable and seldom realistic. While it might be tempting to define $M_k(t)^* = 1 - S_k(t)$ as cause-specific mortality, $M_k(t)^*$ does not obey the conservation of mortality property and generally does not have a meaningful mortality probability interpretation, a point we will return to.

We need to relate $M_k(t) = \Pr(T < t, K = k)$ to the cause-

specific hazard functions $h_k(t)$. Previously, we introduced the cause-specific (conditional) interval mortality as $\Pr(t < T < t + \Delta t, K = k | T > t) = m_k(t + \Delta t | t)$, which we approximated as $\Pr(t < T < t + \Delta t, K = k | T > t) \approx h_k(t)\Delta t$. This is the probability of dying in the interval, conditional on starting the interval alive. To progress, we need the unconditional interval mortality, which is the probability of dying in the interval assuming only that the subject was alive at birth ($T > 0$). Using well-known rules of conditional probability, we can write the unconditional cause-specific interval mortality as $\Pr(t < T < t + \Delta t, K = k | T > 0) \approx h_k(t)\, S(t)\, \Delta t$. The term $h_k(t)\ S(t)$ can be recognized as the probability density function (pdf) corresponding to the cumulative distribution function (cdf) $M_k(t) = \Pr(T < t, K = k)$. Thus, we have

$$M_k(t) = \int_0^t h_k(u) S(u) du \tag{8}$$

This relationship, and not Eq. 6, is the basis for meaningful estimators of cause-specific mortality. In biostatistical survival analysis, researchers call this relationship the cumulative incidence function (CIF; Kalbfleisch and Prentice 2002). Let $f(t)$ be the pdf corresponding to the survival function $S(t)$. An important general relationship in survival analysis that links the hazard function, the survival function, and the pdf is $h(t) = f(t)/S(t)$, or $S(t) = f(t)/h(t)$. This allows us to rewrite the CIF as

$$M_k(t) = \int_0^t r_k(u) f(u) du \tag{9}$$

where $r_k(t)$ is the relative cause-specific hazard defined as $h_k(t)/h(t)$. While Eq. 8 is the form of the CIF most commonly seen in the biostatistical literature, Eq. 9 allows for intuitive insight into the construction of estimators.

Estimators

Before considering cause-specific mortality, we will consider overall mortality. Analogous to Eq. 7, for overall mortality the relationship is

$$M(t) = \int_0^t f(u) du \tag{10}$$

That is, a cdf is obtained by integrating (summing up) the pdf. For now, we will take a strictly nonparametric approach and will not make any assumptions about the functional nature of the hazard function $h(t)$. Recall that $f(t)$ is the pdf of the mortality function $M(t)$. Let $t_i : 0 < t_1 < t_2 < \ldots < t_D$ represent the unique times at which we observed deaths. Intuitively, it seems reasonable not to allow any probability density except at those times t_i when we observe deaths. Because integrals such as Eq. 10 are essentially sums, an estimator of $M(t)$ then becomes a sum describing a step-function:

$$\hat{M}(t_j) = \sum_{i=1}^{j} \hat{f}(t_i) \tag{11}$$

Loosely speaking, $f(t)$ is the probability of surviving from birth until t and then dying, or $f(t) = S(t^-)\ h(t)dt$. $S(t^-)$ means the survival up to a point immediately before time t. As an estimator of $h(t)$, it is reasonable to propose $\hat{h}(t) = d(t)/n(t)$, where $d(t)$ is the number of deaths at instant t, and $n(t)$ is the number of study animals at risk at the same instant. Thus, for $0 < t < t_1$, $\hat{M}(t) = 0$. After t_1 and before t_2, $\hat{M}(t) = \hat{f}(t_1) = d(t_1)/n(t_1)$ [because $\hat{S}(t_1^-) = 1 - \hat{M}(t_1^-) = 1$]. Carrying this a step further for $t_2 < t < t_3$ we have

$$\hat{M}(t) = \hat{f}(t_1) + \hat{f}(t_2) = \frac{d(t_1)}{n(t_1)} + \left[1 - \frac{d(t_1)}{n(t_1)}\right] \frac{d(t_2)}{n(t_2)} \tag{12}$$

because

$$\hat{S}(t_2^-) = \left[1 - \frac{d(t_1)}{n(t_1)}\right]$$

We can apply this process repeatedly for any time t. This procedure is simply the usual GKME given in terms of mortality. The term $\hat{f}(t_i)$ is the height of the step of the GKME at the ith distinct time of death. Each term $\hat{f}(t_i)$ is essentially a parameter, and because the number of such parameters grows as the sample size grows, we refer to such nonparametric approaches as infinite dimensional models, although counterintuitive, nonparametric estimation actually requires estimating the maximum possible number of parameters. As we will discuss, this is less than ideal.

From the GKME, it is now easy to obtain the NPCIFE of cause-specific mortality, based on Eq. 9. In an obvious manner, the NPCIFE becomes

$$\hat{M}_k(t_j) = \sum_{i=1}^{j} \hat{r}_k(t_i) \hat{f}(t_i) \tag{13}$$

The step-size term $\hat{f}(t_i)$ is the same as for the GKME. The obvious estimator for the relative cause-specific hazard is $\hat{r}_k(t) = d_k(t)/d(t)$, where as before $d(t)$ are all deaths at instant t, and $d_k(t)$ are deaths specifically due to cause k. Unless there are tied death times, $\hat{r}_k(t)$ is just a 0/1 indicator function that indicates whether a particular step in the GKME curve should be credited to cause k. In the case of ties, we prorate the step height to the observed causes at time t. Clearly, this approach has the conservation of mortality property with respect to the overall GKME estimator of mortality. Although we presented the NPCIFE based on the KME, the Breslow survival estimator could also be the basis of a NPCIFE and we are aware of no basis for preferring one to the other. Andersen et al. (1993) and Marubini and Valsecchi (2004) describe the rather involved (see Appendix) computation of confidence intervals for $\hat{M}_k(t)$.

We refer to the number of study animals under observation at time t, $n(t)$, as the risk set. In most biomedical applications where staggered entry is not involved, the risk set is largest at time $t = 0$, and declines monotonically with t. Thus in the usual biomedical application of the KME, $M(t)$ is typically well estimated for small t, but small risk sets at larger t can result in

unacceptably high variances in the right tail of $M(t)$. Here we see that things are not quite so simple for staggered entry designs and the GKME. Typically, risk sets for early times may be small or even zero. The GKME becomes undefined if the risk set ever drops to zero. If only a few animals are entered early in the study, deaths occurring early in the study when the risk sets $r(t_i)$ are small will result in high variances for early steps $\hat{f}(t_i)$. Because of the additive structure of $\hat{M}(t)$ and $\hat{M}_k(t)$, the variances will only continue to increase, even if we eventually achieve large risk sets (Pollock et al. 1989).

In such cases, it might be tempting to redefine the origin to some time t^* and ignore animals in the study before time t^*. The obvious problem with this approach is that it achieves no survival estimate for the time interval zero to t^*. A perhaps less obvious problem is that it results in a sampling bias. If we recruited animals before time t^*, such animals still present in the sample by time t^* are biased toward the less frail animals with lower hazards. Thus, the staggered-entry GKME and NPCIFE are less robust than their origin-entry counterparts, and researchers should make every effort to insure that the risk sets are large for early times (DelGiudice et al. 2006).

Compensatory Versus Additive Mortality

The additive nature of hazard functions also provides a useful basis for assessing compensatory versus additive mortality (sensu Bartmann et al. 1992) in wildlife populations. Although hazard functions are mathematically additive, mortality owing to a new (additional) source of mortality is completely additive in the biological sense only if the hazard functions for the existing sources of mortality do not change in response to the addition of each new source of mortality. Compensatory mortality (partial or complete, see Boyce et al. 1999) is inferred when addition (or removal) of a specific hazard to a population changes the original (or remaining) hazard functions. One cannot predict the impact of adding or removing a source of mortality to overall survival probability without knowing how the other hazard functions change in response to this addition or removal.

Proportional Hazard Models

An important property of a hazard function $h(t)$ or cause-specific hazard function $h_k(t)$ is that it remains a hazard function when multiplied by any positive constant. As a counterexample, a probability multiplied by a constant does not necessarily remain a probability if it exceeds 1, which it easily can. Suppose X is some covariate of interest, such as an animal's sex (e.g., $X = 1$ for F, $X = 0$ for M). We could test for a sex effect with the log-linear (multiplicative) model $\log[h(t)] = \log[b(t)] + X\beta$, where β is the sex effect. Here $b(t)$ is the baseline hazard function, which in the example described above is the male hazard function. We can also express this as $h(t) = b(t)\exp(X\beta)$, where $\exp(\beta)$ is the hazard ratio and measures $h(t)/b(t)$. The null hypothesis of no gender difference is equivalent to $\beta = 1$. Although usually not stated as such, this is the hypothesis tested by the usual log-rank test, although the log-rank test does not involve estimating β. Researchers can also test this hypothesis using Cox partial likelihood (proportional hazards approach), which also provides an estimate of β. Proportional hazards models are useful both for comparing cause-specific hazards, and also evaluating how covariates (risk factors) influence hazard functions, both of which we illustrate in our example.

Software for Estimating the NPCIFE

Most standard statistical packages do not contain procedures for computing the NPCIFE or comparing the corresponding differences between groups of animals or causes of mortality. Although various codes for implementing the NPCIFE are available for SPLUS, *R*, STATA, and SAS (e.g., the ***cmprsk*** suite of functions for SPLUS and R, available at http://biowww.dfci.harvard.edu/~gray/, perform the calculations described by Gray [1988]), none allow for left-truncation (staggered entry of animals) and are thus of little applicability for most wildlife studies. However, the computations for the NPCIFE are quite easy given the output from a procedure that computes the GKME. It is essentially a matter of bookkeeping, determining to which cause of mortality to credit each GKME step. Appendix A details the code required to automate the calculation of the NPCIFE, and the associated estimate of variance, using a staggered entry approach in SPLUS. Modifying this code to run in R is straightforward.

Comparing CIFs from Different Causes

Tai et al. (2001) have summarized methods of comparing CIFs. Unfortunately, none of these tests are valid for datasets containing left-truncated data. Therefore, we suggest the following procedures, based on Lunn and McNeil (1995). The approach requires Cox regression software that can accommodate delayed entry data ("counting style input"), such as SAS PROC PHREG or SPLUS coxph(). If there are Q sources of mortality, the data set is duplicated Q times. In the kth data set, all deaths but those due to cause k are censored. One includes $Q - 1$ binary dummy covariates $Z_1, \ldots, Z_{Q-1}$ in the analysis. For the kth data set, $Z_k = 1$ and is zero otherwise. By running models with and without the Zs and performing a likelihood ratio test, one is testing the hypothesis $h_1(t) = h_2(t) \ldots = h_Q(t)$ against the proportional hazards alternative which is that at least one hazard differs by some proportional constant. Like the log-rank test, this test is only sensitive to proportional cause-specific hazards alternatives and would not be sensitive to situations where the hazards cross. Because of this insensitivity to cause-specific nonproportional hazards, we recommend graphic evaluation.

Previously, we defined

$$S_k(t) = \exp\left\{-\int_0^t [h_k(u)]du\right\}$$

and noted that $S_k(t)$ did not have a useful survival probability

Table 1. Example of the data duplication technique discussed by Lunn and McNeil (1995) to perform a stratified Cox's proportional hazards analysis in SPLUS. Data shown are for 50 coyotes radiocollared in Nova Scotia, Canada, during 1992–1996 (Patterson 1999). We replicated each data record for individual marked animals within the same data table, once for each of the causes of mortality under consideration. In this case, we generated separate hazard functions for coyotes lost to harvest-related versus other causes of mortality as indicated by the field "c.strat." The data recorded below indicate that Animal 1 was censored at the end of year 1 (day 365), whereas animal 2 was lost to cause number 1 (harvest) on day 318.[a]

Anim.id	Ent	Exit	Event	Cause	Year	Sex	C.strat
AM1	248	365	0	1	1	M	harvest
AM1	248	365	0	0	1	M	other
AM2	113	318	1	1	1	M	harvest
AM2	113	318	0	0	1	M	other

[a] The full dataset is available http://www3.sympatico.ca/lbpatterson/Data/coydata.zip.

interpretation. $S_k(t)$ is the quantity estimated by the 1-KM procedure, that is, censoring all deaths but those due to k. Although $S_k(t)$ is not useful for estimating probabilities, it provides a useful diagnostic for comparing the cause-specific hazard functions $h_k(t)$. In particular, if we have satisfied the proportional cause-specific hazard assumption, plots of $\log\{-\log[\hat{S}_k(t)]\}$ versus t should produce curves separated by a constant distance reflecting the log hazard ratio between the two hazards. The shapes of both $\log[\hat{S}_k(t)]$ and $\log\{-\log[\hat{S}_k(t)]\}$ are useful for evaluating the functional form of $h_k(t)$. For example, $\log[\hat{S}_k(t)]$ will be linear during intervals where $h_k(t)$ is constant, and $\log\{-\log[\hat{S}_k(t)]\}$ is linear for the class of accelerated failure time models (Kalbfleisch and Prentice 2002).

Evaluating Risk Factors for Cause-Specific Mortality

To allow the comparison of parameter estimates corresponding to different mortality sources, Lunn and McNeil (1995) proposed a simple adaptation of Cox's proportional hazards model. They replicated the entire dataset within a single data table once for each cause of mortality and ran a stratified analysis using cause of death as the stratification factor (e.g., Table 1). The data replication is statistically valid because the hazard functions for each mortality cause are additive with their sum at any time t equaling the overall (all-causes) hazard function (Lunn and McNeil 1995, Therneau and Grambsch 2000). This allows the estimation of separate hazard functions for each cause of mortality and facilitates joint estimation of parameters in a competing risks framework.

Lunn and McNeil (1995) presented methods for both an unstratified and stratified Cox regression. The unstratified Lunn and McNeil method assumes that the hazard functions for different mortality sources are proportional, whereas the stratified method allows for different baseline hazards for each mortality type. In many wildlife studies different mortality sources will be have substantially different underlying hazard functions (e.g., DelGiudice et al. 2002), so we focus on the stratified method in the example discussed below.

Because the hazard functions described above are instantaneous risk functions, they do not quantify the ultimate effect of covariates on the overall magnitude of individual mortality sources (i.e., they do not describe the overall mortality cdf). For example, the instantaneous hazards for multiple mortality sources may cross at various points in time, making visual assessment of likely covariate effects on the cumulative hazards difficult. In response to this concern, Tai et al. (2001) extended the Lunn and McNeil (1995) approach to accommodate the analysis of competing risks based on the NPCIFE. Similarly, Fine and Gray (1999) developed methods for estimating cause-specific CIFs under a semi-parametric log(-log) transformation model. Simulations indicated that the partial likelihood principle and weighting technique they employed yielded valid estimates under a variety of censoring scenarios. They further outlined how to construct confidence intervals based on both analytical and simulation-based techniques. Although potentially appealing to the mathematically astute, the methods proposed by Fine and Gray (1999) are not as easily implemented as the methods of Lunn and McNeil (1995) or Tai et al. (2001), particularly for left-truncated data.

Example and Discussion

We use data from a telemetry study of coyotes in Nova Scotia (Patterson 1999) to illustrate the methods discussed here. We monitored 50 radiocollared coyotes in 2 study areas that differed substantially with respect to prey densities and typical winter severities (Patterson 1999). We prepared data for analysis as in Table 2. The full dataset is available http://www3.sympatico.ca/lbpatterson/Data/coydata.zip. Coyotes entered the analysis from September 1992 to 1996 on the day after we radiocollared them and exited when they died ($n = 28$) or when we censored them on the day of last radio contact ($n = 9$). We censored all coyotes still being monitored when the study terminated on 30 April 1997 ($n = 13$). We assumed the biological year for coyotes extended from 1 May to 30 April and calculated annual estimates of $S(t)$ and $M_k(t)$ by pooling data among years. Most deaths of collared coyotes were due to human related causes (primarily shooting and snaring; Patterson 1999), so we consider here only 2 competing risks: 1) harvest, and 2) other causes.

We calculated cause-specific mortality estimates of $\hat{M}_k(t)^*$ based on the 1-KM method and generated the estimates $\hat{S}_k(t)$ using a null Anderson–Gill proportional hazards model in SPLUS using the following code:

Table 2. Example of the data structure used to estimate $S(t)$, marginal survival distributions based on the 1-KM method (see text), and the NECIF (Appendix) in SPLUS for 50 coyotes radiocollared in Nova Scotia, Canada, during 1992–1996 (Patterson 1999).[a]

Anim.id	Study area	Sex	Age	Year	Ent	Exit	Event	Cause
AM1	1	2	2	2	0	75	1	oth_hum
AM12	3	2	2	5	0	177	1	natural
AM7	3	2	2	4	0	304	1	oth_hum
PM5	3	2	0	3	182	186	1	oth_hum
AF7	2	1	2	4	0	254	1	natural
PF14	2	1	0	4	190	263	1	natural
AF10	2	1	2	4	198	268	1	harvest
AF11	2	1	2	4	208	215	1	harvest
AF12	2	1	2	5	147	192	1	harvest
AF4	3	1	2	3	16	295	1	harvest
AF6	2	1	2	3	166	348	1	harvest
AM11	3	2	2	5	0	325	1	harvest
AM2	2	2	2	3	0	290	1	harvest
AM4	2	2	2	5	0	276	1	harvest
AM5	3	2	1	3	0	290	1	harvest

[a] The full dataset is available http://www3.sympatico.ca/lbpatterson/Data/coydata.zip.

```
Surv.c.1<-coxph(Surv(ent,exit,event)~1,data = file name)
```

where the fields "ent" and "exit" represent the day of entry and exit into the study each year, and "event" was an indicator variable detailing if we censored the animal in question or if it died during the year detailed by each record. We generated the NPCIFE using the code detailed in Appendix A. We calculated seasonal rates by creating subsets of the data containing only exposure days for each collared coyote falling within the respective season.

Annual mortality owing to human-related causes was approximately 24% (Fig. 1) and estimates of harvest-related and other mortality based on the NECIF were slightly lower than those estimated using the 1-KM method (Fig. 1).

Arranging the data as in Table 1 and running the following stratified Cox PH model:

```
<-coxph(Surv(ent,exit,event)~studarea + sex + age +
cluster(id) + strata(c.strat)
```

we tested for the influence of covariates sex, age, and study area on the 2 cause-specific hazards. This analysis suggests that risk of mortality is lowest for adult coyotes (Risk ratio [RR] = 0.52, 95% CI = 0.306 to 0.883), whereas we could not detect a significant influence of study area (RR = 1.33, 95% CI = 0.84 to 2.11) or sex (RR = 2.05, 95% CI = 0.87 to 4.84) on risk of death. More specifically, the approach of Lunn and McNeil (1995) allows for easy estimation of cause-specific coefficients. For example, having determined that age had a significant influence on survival probability, we next asked whether the effect of age was similar for harvested coyotes versus those dying of other causes. We investigated this by coding two dummy variables:

```
> age1<-cysxStr$age*(cysxStr$c.strat == '1')
> age2<-cysxStr$age*(cysxStr$c.strat! = '1')
```

and then running the following model:

```
coxph(Surv(ent,exit,event)~studarea + sex + age1 + age2 +
strata(c.strat)
```

This analysis indicated that age primarily influenced probability of mortality due to harvest (RR = 0.498, 95% CI = 0.304 to 0.816), and had less influence of probability of dying of other causes (RR = 0.621, 95% CI = 0.230 to 1.676).

For comparison, we used the software program MICROMORT (Heisey and Fuller 1985) to implement the HF method of determining cause-specific mortality and survival rates. Because most mortality occurred between October and April each year, we divided the year into 2 seasons for analysis with HF: May–September and October–April. For analysis with the HF method, we tallied the number of radiodays of contact for each interval and recorded all deaths. The results for the NPCIFE and HF were quite similar (Table 3). The smaller confidence intervals associ-

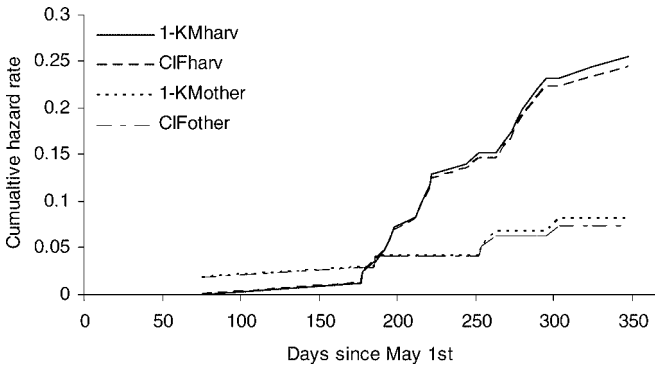

Figure 1. Cumulative probability of mortality owing to harvesting by humans (primarily hunting and trapping) and by other causes (e.g. road-kill, natural mortality, unknown) for 50 coyotes radiocollared in Nova Scotia, Canada, during 1992–1996 (Patterson 1999). For each cause of death, we estimated probability of mortality using both the 1-KM and the NECIF approaches (see text for details).

Table 3. Comparison of overall survival and cumulative hazard rate estimates for mortality owing to harvesting by humans ("harvest"; primarily hunting and trapping) and by other causes ("other"; e.g. road-kill, natural mortality, unknown) for 50 coyotes radiocollared in Nova Scotia, Canada, during 1992–1996 (Patterson 1999). For each cause of death, we estimated cumulative hazards using both the 1-KM and the NECIF approaches (see text for details).[a]

	May–Sep		Oct–Apr		Annual	
	Rate	95% CI	Rate	95% CI	Rate	95% CI
Heisey–Fuller method						
$S(t)$	0.961	0.909, 1.0	0.668	0.572, 0.780	0.642	0.544, 0.757
m_{kt} (harvest)			0.281	0.182, 0.380	0.270	0.174, 0.366
m_{kt} (other)	0.039	0, 0.092	0.051	0.002, 0.100	0.088	0.019, 0.157
NPCIFE						
$S(t)$	0.965	0.925, 0.995	0.714	0.608, 0.780	0.679	0.617, 0.744
m_{kt} (harvest)			0.242	0.168, 0.316	0.242	0.168, 0.316
m_{kt} (other)	0.035	0.005, 0.075	0.044	0.002, 0.086	0.079	0.027, 0.131

[a] The full dataset is available http://www3.sympatico.ca/lbpatterson/Data/coydata.zip.

ated with the NPCIFE are somewhat puzzling and probably reflect slight differences in how we performed the variance approximations. Simulations would be useful to examine the accuracy of the confidence interval coverages.

Given the relatively close correspondence between the estimates produced by the 1-KM and NECIF estimators, one question that naturally arises is whether the bias associated with the 1-KM estimator of cause-specific mortality is serious enough to be of practical concern. We can gain insight into this by considering a simple case where hazards are constant over time (or age). Assume the hazards for two sources of mortality are h_1 and h_2, so the total hazard is $h = h_1 + h_2$ and the relative hazard (i.e. proportion of deaths expected) due to cause 1 is $r_1 = h_1/h$. Overall survival is $S(t) = \exp(-ht)$. The NPCIFE for cause 1 is a consistent estimator of $M_1(t) = r_1[1 - S(t)]$. Clearly, as time t grows large, this quantity approaches the relative hazard due to cause 1, r_1. 1-KM is a consistent estimator of $M^*_1(t) = 1 - \exp(h_1 t)$. We cannot give this any reasonable definition as an observable probability (Kalbfleisch and Prentice 2002). As t grows large, this quantity approaches 1. Note that this quantity can also be expressed as $M^*_1(t) = 1 - S(t)\exp(-r_1 t)$. The ratio of $M^*_1(t)/ M_1(t)$ defines the expected relative bias $\beta(t)$ of 1-KM relative to the NECIF, and we can express it solely as a function of time t and the relative hazard r_1:

$$\beta\ (t) = \frac{1 - S(t)\exp(-r_1 t)}{r_1[1 - S(t)]}$$

The expected relative bias for relative hazards of $\theta = 0.20$, $\theta = 0.50$, and $\theta = 0.70$ is shown in Fig. 2. In all cases, as t grows large, the relative expected bias $\beta(t)$ approaches $1/r_1$. This illustrates several general points. First, assuming constant hazards, the relative bias of 1-KM relative to the NPCIFE increases as the total mortality increases. Furthermore, the relative bias of 1-KM versus the NPCIFE has an inverse relationship to the relative contribution of the mortality source in question. That is, relative bias will be greatest for mortality sources that represent only a small fraction of overall mortality. Most importantly, these biases are non-negligible for a wide range of biologically realistic total mortalities and relative hazards (Fig. 2). This underscores the statement by Gooley et al. (1999) that "Determining whether to use 1-KM or CIF is unambiguous and CIF should always be used if an estimate of the probability of failure of a specific type is desired."

As noted, both the NPCIFE and HF have their shortcomings. Especially when applied to staggered entry data, the NPCIFE can be overparameterized. The result is that in some cases, the NPCIFE may have unacceptable high variances or even be undefined. On the other hand, HF requires arbitrary interval construction, which can lead to biases if misplaced. As mentioned previously, the single best criterion of an estimator is the MSE, which balances variance and bias. Compromise models, models that are smoother than strictly nonparametic models, but not "too" smooth, typically achieve optimal MSEs. In survival analysis, a popular approach is to model the hazard as a spline function; in fact, the piecewise HF model is a simplistic spline hazard model (a 0-order b-spline). Spline models can be of 2 types, smoothing splines and regression splines. Regression spline models require the arbitrary placement of knots; HF is a regression spline model. Smoothing spline models do not require knot placement but are generally computationally more demanding. Penalized regression spline (PRS) models provide a useful compromise. Although the researcher must still decide on knot placement for PRS models, parsimony is not an issue and the researcher should initially specify more knots than are likely to be necessary. Researchers prevent over-parameterization by adding a roughness penalty term in the log likelihood. Then, using a criterion such as AIC (e.g., Heisey and Foong 1998), researchers apply an optimal smoothing factor to the spline approximation to prevent wild variation of the hazard due simply to sparse data. This approach is similar to Bayesian approaches; the penalty acts in much the same manner as would a prior distribution imposed on the hazard function. We believe such regularization approaches

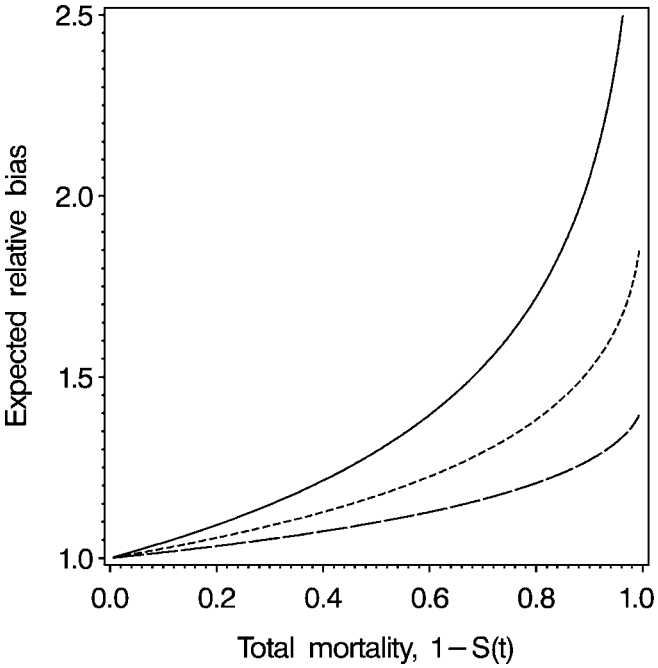

Figure 2. The relative bias of 1-KM compared to the NECIF. We show 3 values of the relative hazard: $\theta = 0.20$ (solid line), $\theta = 0.50$ (small-dashed line), and $\theta = 0.70$ (large-dashed line). We explain the vertical lines in the text.

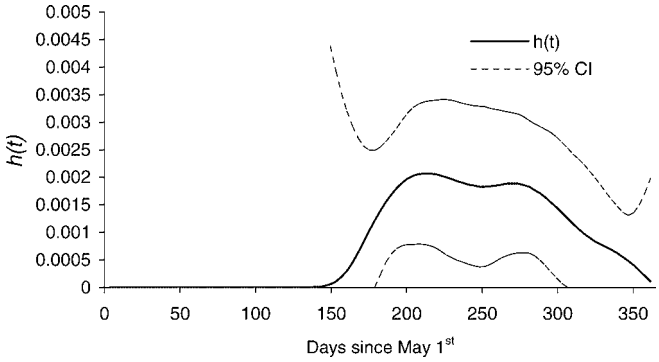

Figure 3. Seasonal changes in the penalized likelihood estimate (PLE) of the harvest-related hazard for 50 coyotes radiocollared in Nova Scotia, Canada, during 1992–1996 (Patterson 1999).

arguably offer the most promising future direction for CSM analyses.

We estimated the penalized likelihood estimate (PLE) of the instantaneous human-causes hazard, $h_k(t)$, using program PHMPL (Joly et al. 1998, 1999) to illustrate seasonal changes in the instantaneous risk of coyotes being killed by humans. We assigned 17 knots and allowed PHMPL to determine the smoothing parameter automatically (Fig. 3; Joly et al. 1998, 1999). The instantaneous hazard could also have been easily estimated using a kernel-based approach implemented by the SPLUS code "muhaz" (Mueller and Wang 1994), available for download at http://odin.mdacc.tmc.edu. For brevity, we do not present the resulting cause-specific mortality estimates, but they are relatively easy to obtain via Eq. 4 and 8. Bootstrapping is the best method to obtain variances.

Management Implications

The statistical issues involved in cause-specific mortality estimation can be rather complex, and intuition does not always serve one in good stead as exemplified by the frequent misuse of the 1-KM estimator. Survival analysis remains a relatively specialized domain with a human biomedical emphasis, and most survival analyses courses in biostatistics or public health departments teach from that perspective. From a statistical standpoint, the issues of survival analysis such as truncation and censoring are unique, and many otherwise well-trained statisticians may have little exposure to these techniques. Wildlife researchers working with survival data should seek out the assistance of biostatistical survival analysis experts; some will welcome the novelty of nonhuman applications. Wildlife graduate students should be encouraged to pursue biostatistical survival analysis coursework, even if the focus is on human applications.

In the meantime, managers will continue to require good estimates of the magnitude of various mortality sources for a variety of wildlife species, and the NPCIFE will often meet that need. The NPCIFE is simple, yet not widely implemented, probably because of a lack of readily available software for performing the required calculations. A drawback of using CIFs to estimate cause-specific mortality functions is that they do not treat different causes of mortality jointly, thus complicating comparison of parameter estimates corresponding to different mortality sources. However, an attractive feature of the NPCIFE is that it meshes naturally with proportional hazards models such as those presented by Lunn and McNeil (1995) to allow assessment of relative impact of a covariate(s) on various mortality sources (e.g. assessing the impact of forest harvesting on losses of moose to hunters vs. natural predators).

Acknowledgments

We thank K. Middel, Ontario Ministry of Natural Resources, for writing the code to generate the cumulative incidence function in SPLUS. J. Bowman, T. MacDonald, D. L. Murray, and G. C. White provided useful comments on a previous draft of this manuscript.

Literature Cited

Andersen, P. K., O. Borgan, R. D. Gill, and N. Keiding. 1993. Statistical models based on counting processes. Springer-Verlag, New York, New York, USA.

Banfield, A. W. F. 1955. A provisional life table for the barren ground caribou. Canadian Journal of Zoology 33:143–147.

Bartmann, R. M., G. C. White, and L. H. Carpenter. 1992.

Compensatory mortality in a Colorado mule deer population. Wildlife Monographs 121:1–39.
Boyce, M. S., A. R. E. Sinclair, and G. C. White. 1999. Seasonal compensation of predation and harvesting. Oikos 87:419–426.
Cai, T. X., and R. A. Betensky. 2003. Hazard regression for interval-censored data with penalized spline. Biometrics 59:570–579.
Caughley, G. 1966. Mortality patterns in mammals. Ecology 47:906–918.
Cox, D. R. 1972. Regression models and life tables (with discussion). Journal of the Royal Statistical Society, Series B 34:187–220.
Cox, D. R., and D. Oakes. 1984. Analysis of survival data. Chapman and Hall, London, United Kingdom.
DelGiudice, G. D., J. Fieberg, M. R. Riggs, M. Carstensen-Powell, and W. Pan. 2006. A long-term age-specific survival analysis of female white-tailed deer. Journal of Wildlife Management 70:1556–1568.
DelGiudice, G. D., M. R. Riggs, P. Joly, and W. Pan. 2002. Winter severity, survival, and cause-specific mortality of female white-tailed deer in north-central Minnesota. Journal of Wildlife Management 66: 698–717.
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.
Gaynor, J. J., E. J. Feuer, C. C. Tan, D. H. Wu, C. R. Little, D. J. Straus, B. D. Clarkson, and M. F. Brennan. 1993. On the use of cause-specific failure and conditional failure probabilities: examples from clinical oncology data. Journal of the American Statistical Association 88:400–409.
Gill, R., and S. Johansen. 1990. A survey of product-integration with a view towards application in survival analysis. Annals of Statistics 18: 1501–1555.
Gooley, T. A., W. Leisenring, J. Crowley, and B. E. Storer. 1999. Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. Statistics in Medicine 18:695–706.
Gray, R. 1988. A class of *k*-sample test for comparing the cumulative incidence of a competing risk. Annals of Statistics 16:1141–1154.
Heisey, D. M., and A. P. Foong. 1998. Modeling time-dependent interaction in a time-varying covariate and its application to rejection episodes and kidney transplant failure. Biometrics 54:712–719.
Heisey, D. M., and T. K. Fuller. 1985. Evaluation of survival and cause-specific mortality rate using telemetry data. Journal of Wildlife Management 49:668–674.
Heisey, D. M., T. L. Shaffer, and G. C. White. 2006. In press. The ABC's of nest survival: theory and application from a biostatistical perspective. Studies in Avian Biology.
Joly, P., D. Commenges, and L. Letenneur. 1998. A penalized likelihood approach for arbitrarily censored and truncated data: application to age-specific incidence of dementia. Biometrics 54: 185–194.
Joly, P., L. Letenneur, A. Alioum, and D. Commenges. 1999. PHMPL: a computer program for hazard estimation using a penalized likelihood method with interval-censored and left-truncated data. Computer Methods and Programs in Biomedicine 60:225–231.
Kalbfleisch, J. D., and R. L. Prentice. 2002. The statistical analysis of failure time data, 2nd edition. Wiley, New York, New York, USA.
Kaplan, E. L., and P. Meier. 1958. Non-parametric estimation from incomplete observations. Journal of the American Statistical Association 53:457–481.
Kurzejeski, L., D. Vangilder, and J. B. Lewis. 1987. Survival of wild turkey hens in northern Missouri. Journal of Wildlife Managment 51: 188–193.
Laird, N., and D. Oliver. 1980. Covariance analysis of censored survival data using log-linear analysis techniques. Journal of the American Statistical Association 90:708–716.
Lindgren, B. W. 1976. Statistical theory. Third edition. Macmillan, New York, New York, USA.
Lunn, M., and D. McNeil. 1995. Applying Cox regression to competing risks. Biometrics 51:524–532.
Lynden-Bell, D. 1971. A method of allowing for known observational selection in small samples applied to 3CR quasars. Monthly Notices of the Royal Astronomical Society 155:95–118.
Marubini, E., and M. G. Valsecchi. 2004. Analyzing survival data from clinical trials and observational studies. John Wiley and Sons, New York, New York, USA.
Mueller, H. G., and J. L. Wang. 1994. Hazard rates estimation under random censoring with varying kernels and bandwidths. Biometrics 50:61–76.
Murray, D. L. 2006. On improving telemetry-based survival estimation. Journal of Wildlife Management 70:1530–1543.
Patterson, B. R. 1999. The effects of prey distribution and abundance on eastern coyote life history and predation on white-tailed deer. Dissertation, University of Saskatchewan, Saskatoon, Canada.
Pepe, M. S., and M. Mori. 1993. Kaplan–Meier, marginal or conditional probability curves in summarizing competing risks failure time data? Statistics in Medicine 12:737–751.
Pollock, K. H., S. R. Winterstein, and C. M. Bunck. 1989. Survival analysis in telemetry studies: the staggered entry design. Journal of Wildlife Management 53:7–15.
Tai, B. C., D. Machin, I. White, and V. Gebski. 2001. Competing risks analysis of patients with osteosarcoma: A comparison of four different approaches. Statistics in Medicine 20:661–684.
Therneau, T. M., and P. M. Grambsch. 2000. Modeling survival data: extending the Cox model. Springer-Verlag, New York, New York, USA.
Trent, T. T., and O. J. Rongstad. 1974. Home range and survival of cottontail rabbits in southwestern Wisconsin. Journal of Wildlife Management 38:459–472.
Tsai, K., K. H. Pollock, and C. Brownie. 1999. Effects of violation of assumptions for survival analysis methods in radiotelemetry studies. Journal of Wildlife Management 63:1369–1375.
Tsai, W. Y., N. P. Jewell, and M. C. Wang. 1987. A note on the product-limit estimator under right-censoring and left truncation. Biometrika 74:883–886.
Williams, B. K., J. D. Nichols, and M. J. Conroy. 2002. Analysis and management of animal populations. Academic, New York, New York, USA.
Winterstein, S. R., K. P. Pollock, and C. M. Bunck. 2001. Analysis of survival data from radiotelemetry studies. Pages 351–380 *in* J. J. Millspaugh and J. M. Marzluff, editors. Radio tracking and animal populations. Academic, New York, New York, USA.

Associate Editor: *White*.

Appendix. Calculating cause-specific cumulative incidence functions in SPLUS.

```
# Script for creating table of cumulative
  survival
# run the following script, then attach the
  table of survival to the current
# data frame. The fields must have the names
  "ent", "exit", "event" and "cause"
# all in lower case letters. Once the script
  is run, there will be a function in your
  current chapter
# cause.survival. Usage of this function is
  cause.survival(table, causenumber).
# Where table is the table with the data,
  causenumber is an integer value used to
  identify the cause of mortality.
"cause.survival.v2" = function(table, p)
{
  assign("p", p, frame = 1)
  # create tables to hold the results of GKM
  survival estimates for all events and for
  cause specific event
  temp.all <- summary(survfit(Surv(ent,
  exit, event) ~ 1, data = table))
```

```
 temp.s <- summary(survfit(Surv(ent, exit,
 cause == p) ~ 1, data = table))
 #combine the two tables so survival of all
 events can be combined with those of the
 cause specific events
 s.df <- data.frame(time = temp.s$time,
 n.event = temp.s$n.event, n.risk =
 temp.s$n.risk, survival = temp.s$surv)
 all.df <- data.frame(time =
 temp.all$time, n.event =
 temp.all$n.event, n.risk =
 temp.all$n.risk, survival =
 temp.all$surv)
 all.s.df <- merge(all.df, s.df, by.x =
 "time", by.y = "time", all.x = T, suffixes =
 c(".all", ".s"))
 assign("n", all.s.df, frame = 1)
 x <- numRows(n)
 #create temporary placeholders for the
 calculation of the mortality rate and the
 cause-specific cumulative incidence func.
 tmp.string <- numeric(x)
 tmp.string2 <- numeric(x)
 t <- 1
 #cycle through the records of the table,
 including all events to calculate
 mortality rate and CIF
 while(t <= x) {
 tmp.string[1] <- n$n.event.s[1]/
 n$n.risk.s[1]
 tmp.string[t] <- (n$survival.all[t - 1] *
 n$n.event.s[t])/n$n.risk.s[t]
 if(tmp.string[t] == "NA") tmp.string2[t]
 <- NA else tmp.string2[t] <-
 sum(tmp.string[1:t], na.rm = T)
 t = t + 1
 }
 MORT <- data.frame(mort.rate =
 tmp.string)
 CIF2 <- data.frame(CIF = tmp.string2)
 CIF.s.all <- cbind(all.s.df, MORT, CIF2)
# Calculate the variance, standard error
and the Confidence Intervals around CIF
SE <- numeric(x)
totvar.t <- numeric(x)
#Reset all temporary variables
t <- 1
j <- 1
Ij <- 0
cumvar.p1 <- 0
cumvar.p2 <- 0
cumvar.p3 <- 0
#loop for the total number of records
while (t <= x)
{
 It <- CIF.s.all$CIF[t]
 if(It == "NA") {
 CIF.s.all$cumvar[t] <- "NA"
 CIF.s.all$StdErr[t] <- "NA"
 CIF.s.all$CI.u[t] <- "NA"
 CIF.s.all$CI.l[t] <- "NA"
 t = t + 1
}
else
{
 while (j < t)
  {
 if(CIF.s.all$CIF[j] == "NA")
  Ij <- Ij
 else
  Ij <- CIF.s.all$CIF[j]
 cumvar.p1 <- cumvar.p1 + (It - Ij)^2 *
(CIF.s.all$n.event.all[j] /
(CIF.s.all$n.risk.all[j] *
(CIF.s.all$n.risk.all[j] -
CIF.s.all$n.event.all[j])))
 if(CIF.s.all$CIF[j] != "NA")
 {
  if(j == 1)
   Sj3 <- 1
  else
  Sj3 <- CIF.s.all$survival.all[j-1]
  Ijc <- CIF.s.all$CIF[j]
  cumvar.p3 <- cumvar.p3 + (It -
Ijc)*(Sj3)*(CIF.s.all$n.event.all[j] /
(CIF.s.all$n.risk.all[j])^2)
 }
 j <- j + 1
 }
 if (t == 1)
  Sj2 <- 1
 else
  Sj2 <- CIF.s.all$survival.all[t-1]
 cumvar.p2 <- (Sj2)^2 *
(((CIF.s.all$n.event.all[t])*(CIF.
s.all$n.risk.all[t] -
CIF.s.all$n.event.all[t]))/
(CIF.s.all$n.risk.all[t])^3) + cumvar.p2
 #total all three components of the
 variance equation to get the final
 variance, generate std. err and
 confidence intervals
 #Assign all results to the output table
 totvar.t[t] <- cumvar.p1 + cumvar.p2 - (2
 * cumvar.p3)
 CIF.s.all$cumvar[t] <- totvar.t[t]
 SE[t] <- sqrt(totvar.t[t])
 CIF.s.all$StdErr[t] <- SE[t]
 CIF.s.all$CI.u[t] <- CIF.s.all$CIF[t] +
 (1.64 * SE[t])
 CIF.s.all$CI.l[t] <- CIF.s.all$CIF[t] -
 (1.64 * SE[t])
 t = t + 1
 j <- 1
```

```
  }
cumvar.p1 <- 0
cumvar.p3 <- 0
Ij <- 0
It <- 0
}
#Variance calculations end here ————————
  return(CIF.s.all)
}
```