STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:1170–1180
Published online 10 August 2006 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2661

Comparing the small sample performance of several variance estimators under competing risks

Thomas M. Braun*, † and Zheng Yuan‡

Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, U.S.A.

SUMMARY

We examine several variance estimators for cumulative incidence estimators that have been proposed over time, some of which are derived from asymptotic martingale or counting process theory, and some of which are developed from the moments of the multinomial distribution. There is little published work comparing these variance estimators, largely because the variance estimators are algebraically complex and difficult to interpret and all but one have yet to be programmed for a standard statistical package. Through simulation and application to real data, we compare the performance of six variance estimators in relation to each other and the bootstrap in order to confirm earlier reports of their performance and to provide future direction toward their application. We find that the multinomial-moment-based estimators have performance close to that of the bootstrap, and are quite accurate for estimating the variance, even in samples of 20 subjects. All but one of the martingale theory-based estimators tend to perform poorly in small samples, tending to either overestimate or underestimate the empirical variance in samples of fewer than 100 subjects. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: cumulative incidence; multinomial; counting process; bootstrap

1. INTRODUCTION

The work of Kaplan and Meier (KM) revolutionized non-parametric analysis of time-to-event clinical outcomes in situations when the outcome of interest is not observed in some subjects, either through loss-to-follow-up or administrative censoring [1]. However, in the presence of competing risks, i.e. the event of interest could be preceded by another event, use of the KM estimator is problematic. Specifically, the KM estimator for each event, treating the competing event as censoring, estimates the marginal survival function only if the events are independent [2, 3]. However, one cannot use the data collected to verify independence of the events because

*Correspondence to: Thomas M. Braun, 1420 Washington Heights, M4063 SPH II, Ann Arbor, MI 48109, U.S.A.
†E-mail: tombraun@umich.edu
‡E-mail: yuanz@umich.edu

Contract/grant sponsor: National Institutes of Health; contract/grant number: 5 P30 CA46592

Received 26 October 2005
Accepted 13 June 2006

every joint distribution of two dependent competing events can be represented by a joint distribution of two independent events [2]. As a slight remedy to this non-identifiability, one can use the data to develop bounds for the true marginal survival distributions [3, 4].

Our motivating example lies in the research of allogenic bone marrow transplants (alloTx), in which estimating the probability of relapse is complicated by the prevalence of acute graft-*versus*-host-disease (aGHVD), an outcome which often precedes relapse and is fatal if left untreated. However, treatment of aGVHD often modulates the graft-*versus*-leukaemia (GVL) effect, and thus, the underlying probability of relapse [5]. As a result, when assessing naturally occurring relapses (i.e. those not due to unanticipated side-effects of medical intervention) in alloTx patients, aGVHD would be a competing risk. Similarly, death due to causes other than aGVHD or relapse, such as infection or other transplant complications, would be a competing event for both relapse and aGVHD. Other clinical settings in which investigators have examined the effect of competing risks include osteosarcoma [6], hepatitis C [7], heart valve replacement [8], organ transplantation [9], and the effects of radiation on normal tissue in cancer patients [10].

Continuing with our example, we first assume that aGVHD and relapse cannot occur simultaneously and we observe only the earlier of the two events in non-censored subjects. Assuming that occurrences of relapse and aGVHD are independent, one could use the KM estimator to estimate $1 - S_r(t) = 1 - \exp\{-\int_0^t \lambda_r(u)\,\mathrm{d}u\}$, where $\lambda_r(u)$ is the relapse-specific hazard rate, and subjects who develop aGVHD or die prior to relapse would be censored for relapse at the time of the competing event. If one were interested instead in the probability of relapse without assuming independence, the relevant quantity is $\mathrm{CI}_r(t) = \int_0^t S(u)\lambda_r(u)\,\mathrm{d}u$, in which $S(u)$ is the survivor function for the earliest of relapse, aGVHD and death. Although most commonly referred to as cumulative incidence, $\mathrm{CI}_r(t)$ has also been referred to as the cause-specific risk [11], the crude incidence curve [12], and the cause-specific failure probability [13]. It is also well-established that $\mathrm{CI}_r(t) \leqslant 1 - S_r(t) \leqslant 1 - S(t)$ [4].

In many cancer-related settings, $\mathrm{CI}_r(t)$ is an observable and seemingly more relevant quantity than $1 - S_r(t)$, although the preference of $\mathrm{CI}_r(t)$ to $1 - S_r(t)$ is subject to debate. For example, although Caplan *et al.* [10] promoted the use of $\mathrm{CI}_r(t)$ when reporting the late normal-tissue effects, i.e. toxicity, of radiation, Bentzen *et al.* [14] countered with an argument for using $1 - S_r(t)$. An excellent summary of the debate was later summarized by Chappell [15]. Bentzen *et al.* argued that cumulative incidence for an event is difficult to interpret on its own because it is dependent upon the incidence of the competing event. Specifically, one can eliminate the cumulative incidence of late normal-tissue effects by insuring that the toxicity is always preceded by death. Put another way, reporting a cumulative incidence to patients is not fully informative as cumulative incidence blends together the experience of subjects who died before toxicity with subjects whose normal tissue was truly not affected. See Farley *et al.* [16] for a discussion of this debate in the setting of time-to-discontinuation of inter-uterine devices (IUDs).

Some authors have also presented situations with the undesirable property $\sum_r [1 - S_r(t)] > 1 - S(t)$, further complicating the use of the KM estimator with competing risks [13, 17]. As a result, statisticians generally recommend using $\mathrm{CI}_r(t)$ instead of $1 - S_r(t)$, although as suggested by Pepe and Mori [17], the cumulative incidence of one event must be interpreted in relation to the cumulative incidence of its competing events. In our example, one should report the cumulative incidences of both aGVHD and relapse, as a low cumulative incidence of relapse would not be interpreted favourably if accompanied by a high cumulative incidence of aGVHD.

The cumulative incidence estimator $\widehat{\mathrm{CI}}_r(t) = \int_0^t \widehat{S}(u)\widehat{\lambda}_r(u)\,\mathrm{d}u$, is simply a cumulative sum of the KM estimator at time t for the earliest of relapse, aGVHD, and death weighted by the observed

DOI: 10.1002/sim

proportion of subjects who relapse by time t. The fundamental theoretical work of the cumulative incidence estimator and its variance was first described by Aalen [18] and Aalen and Johansen [19] using a non-homogeneous Markov process formulation, which is also described in Chapter 4 of the text by Andersen *et al.* [20]. Although more recent work on $\widehat{\mathrm{CI}}_r(t)$ has been typically based in martingale and counting-process theory, Gooley *et al.* [21] contains an elegant derivation of $\widehat{\mathrm{CI}}(t)$, showing how it is related to the KM estimator, $\mathrm{KM}_r(t) = \widehat{S}_r(t)$ and how $1-\mathrm{KM}_r(t) \geqslant \widehat{\mathrm{CI}}_r(t)$, with equality only in the absence of competing risks. Furthermore, Betensky and Schoenfeld [22] published work showing the relationship of estimating cumulative incidence and estimating the cure fraction in cure models.

However, there is little published work comparing some of the existing variance estimators for cumulative incidence estimates [12, 13, 22–25]. A primary reason for this dearth of research is that all of the variance estimators are algebraically complex and difficult to interpret, although Gray's work [23] has been programmed and documented for general use (via the *cuminc* package on the Comprehensive *R* Archive Network, http://cran.r-project.org). Furthermore, it remains unclear as to which variance estimator to use in practical applications. For example, Betensky and Schoenfeld report that Gray's method tends to overestimate the true variance [22], while Pepe reports that her method tends to underestimate the true variance [24]. We have also discovered through our own work that the variance estimator of Betensky and Schoenfeld is algebraically equivalent to that of Gaynor *et al.*

A thorough examination of these six variance estimators is needed and is the primary motivation of this manuscript. Section 2 describes each of the six variance estimators, while Section 3 examines the performance of each of the estimators in a variety of numerical examples and an actual setting. Concluding remarks are found in Section 4.

2. DESCRIPTION OF VARIANCE ESTIMATORS

2.1. Approaches based on counting process/martingale theory

Suppose we have a study of N subjects and subject $i, i = 1, 2, \ldots, N$ can experience each of J competing events, one of which is being censored. Theoretically, subject i will experience event j, $j = 1, 2, \ldots J$, at time T_{ij}, and we observe $X_i = \min\{T_{i1}, T_{i2}, \ldots, T_{iJ}\}$ with an indicator $\delta_i = j$ if $X_i = T_{ij}$. $\mathrm{CI}_j(t)$ is the cumulative incidence of event j at time t, with corresponding estimate $\widehat{\mathrm{CI}}_j(t)$. Note that we assume censoring is independent of all other events in order to insure that $\widehat{\mathrm{CI}}_j(t)$ is consistent for $\mathrm{CI}_j(t)$.

We define $Y_i(t) = I(X_i \geqslant t)$ as the indicator that subject i was at risk at time t for all J events and $N_{ij}(t) = I(X_i \leqslant t, \delta_i = j)$ as the indicator that subject i experienced event j by time t. The total number of subjects at risk at time t is $\bar{Y}(t) = \sum_i Y_i(t)$, and the number of subjects experiencing event j by time t is $\bar{N}_j(t) = \sum_i N_{ij}(t)$. We define the martingale $M_{ij}(t) = N_{ij}(t) - \int_0^t Y_i(u)\lambda_j(u)$, where $\lambda_j(u)$ is the hazard associated with event j.

Lin [25] shows that $M_{ij}(t)$ is asymptotically equivalent to $G_{ij}N_{ij}(t)$, where G_{ij} are independent standard normal variables and that $W_j(t) = n^{1/2}[\widehat{\mathrm{CI}}_j(t) - \mathrm{CI}_j(t)]$ is asymptotically equivalent to $n^{1/2}\sum_{i=1}^{N}[W_{ij1}(t) + W_{ij2}(t) - W_{ij3}(t)]$ where

$$W_{ij1}(t) = \int_0^t \frac{[1 - \sum_{k=1;k\neq j}^{J} \widehat{\mathrm{CI}}_k(t)]}{\bar{Y}(u)} G_{ij}\, \mathrm{d}N_{ij}(u) \tag{1}$$

DOI: 10.1002/sim

$$W_{ij2}(t) = \int_0^t \frac{\widehat{\mathrm{CI}}_j(u)}{\bar{Y}(u)} \sum_{k=1;k\neq j}^{J} G_{ik}\,\mathrm{d}N_{ik}(u) \tag{2}$$

$$W_{ij3}(t) = \widehat{\mathrm{CI}}_j(t) \int_0^t \frac{1}{\bar{Y}(u)} \sum_{k=1}^{J} G_{ik}\,\mathrm{d}N_{ik}(u) \tag{3}$$

Thus, we generate B samples from a standard normal distribution for each G_{ij}, where B is a reasonable number of samples to generate valid results (Lin uses $B = 500$ in his manuscript). We then combine the sampled G_{ij} with our observed values of $\{Y_i(t), N_{i1}(t), \ldots, N_{iJ}(t)\}$ in equations (1)–(3) to create B values of $W_j(t)$, and the sample variance of those B values divided by n gives us our estimate for the variance of $\widehat{\mathrm{CI}}_j(t)$.

Pepe [24] shows that $W_j(t)$ is asymptotically equivalent to $n^{1/2}\sum_i^N V_{ij}(t)$ where the $V_{ij}(t) = \sum_{i=1}^N [V_{ij1}(t) - V_{ij2}(t)]$ are independent, mean-zero variables in which

$$V_{ij1}(t) = \int_0^t \frac{\widehat{S}_*(u)}{\bar{Y}(u)}[\mathrm{d}N_{ij}(u) - Y_i(u)\hat{\lambda}_j(u)\,\mathrm{d}u]$$

$$V_{ij2}(t) = \int_0^t \widehat{S}_*(u)\left\{\int_0^u \frac{1}{\bar{Y}(v)}[\mathrm{d}N_{i*}(v) - Y_i(v)\hat{\lambda}_*(v)]\,\mathrm{d}v\right\}\hat{\lambda}_j(u)\,\mathrm{d}u$$

in which $\hat{\lambda}_j(u) = \bar{N}_j(u)/\bar{Y}(u)$. The subscript asterisk indicates a function for the earliest of all J events, so that $\widehat{S}_*(u)$ is simply the Kaplan–Meier estimate for the probability of surviving all J events and $\hat{\lambda}_*(u) = \sum_j \bar{N}_j(u)/\bar{Y}(u)$ its corresponding hazard estimate. Therefore, $\sum_i V_{ij}^2(t)/n$ gives us an estimate for the variance of $\widehat{\mathrm{CI}}_j(t)$.

Korn and Dorey [12] propose a third variance estimator by taking the asymptotic variance derived by Aalen [18] and plugging in consistent estimators for the unknown values in the estimator, while a fourth variance estimator has foundations in Gray's landmark paper for comparing two or more cumulative incidence curves [23]. We omit further details of these two estimators due to space limitations.

2.2. *Approaches based on multinomial distribution*

At any time t, we have $\bar{Y}(t)$ subjects at risk for all J events prior to time t, and $\bar{N}_j(t)$ subjects who have since experienced event j. Therefore, $\bar{\mathbf{N}}(t) = \{\bar{N}_1(t), \bar{N}_2(t), \ldots, \bar{N}_J(t)\}$ has a multinomial distribution with parameters $\bar{Y}(t)$ and $\boldsymbol{\lambda}(t) = \{\lambda_1(t), \lambda_2(t), \ldots, \lambda_J(t)\}$.

If we take the N observed times of follow-up $\{X_1, X_2, \ldots, X_N\}$ and identify the $M \leqslant N$ unique ordered event times $\{Z_1, Z_2, \ldots, Z_M\}$, then we have

$$\widehat{\mathrm{CI}}_j(t) = \sum_{Z_i \leqslant t} \hat{\lambda}_j(Z_i)\left\{\prod_{Z_i < t}[1 - \sum_j \hat{\lambda}_j(Z_i)]\right\} \tag{4}$$

$$= \sum_{Z_i \leqslant t} \hat{\lambda}_j(Z_i)\widehat{S}(Z_i) \tag{5}$$

DOI: 10.1002/sim

so that

$$\mathrm{Var}\{\widehat{\mathrm{CI}}_j(t)\} = \sum_{Z_i \leqslant t} \mathrm{Var}\{\hat{\lambda}_j(Z_i)\widehat{S}(Z_i)\} + 2\sum_{Z_i \leqslant t}\sum_{Z_i < Z_{i'} \leqslant t} \mathrm{Cov}\{\hat{\lambda}_j(Z_i)\widehat{S}(Z_i), \hat{\lambda}_j(Z_{i'})\widehat{S}(Z_{i'})\} \quad (6)$$

As the multinomial distribution tells us that

$$E[\hat{\lambda}_j(t)] = \lambda_j(t)$$

$$\mathrm{Var}[\hat{\lambda}_j(t)] = \lambda_j(t)[1 - \lambda_j(t)]/\bar{Y}(t)$$

$$\mathrm{Cov}\{\hat{\lambda}_j(t), \hat{\lambda}_k(t)\} = -\lambda_j(t)\lambda_k(t)/\bar{Y}(t), \quad j \neq k$$

we can estimate the variance of $\widehat{\mathrm{CI}}_j(t)$ once we estimate the moments of $\hat{\boldsymbol{\lambda}}(t)$, which although straightforward, is a complicated task due to the recursive nature of equations (4) and (5).

Betensky and Schoenfeld [22] cite the approximate independence of $\hat{\lambda}_j(u)\widehat{S}(u)$ and $\hat{\lambda}_j(v)\widehat{S}(v)$, $u \neq v$, thereby ignoring the covariance terms in equation (6), and describe how the remaining variance terms in equation (6) can be computed. Gaynor *et al.* [13] took a seemingly different approach by using single-order Taylor series approximations for the variance and covariance terms in equation (6), which is similar to Greenwood's derivation of the variance of the KM estimator [26]. However, the approaches Betensky and Schoenfeld and Gaynor both have foundations in the work of Dinse and Larson [27], and with some algebra, one can show that both approaches are equivalent to each other. Essentially, the omission of higher order terms by Gaynor *et al.* is equivalent to the assumption by Betensky and Schoenfeld of independence of estimates at different timepoints. Choudhury [28] also applied the methods of Dinse and Larson to the computation of confidence intervals for cumulative incidence estimators.

3. OPERATING CHARACTERISTICS

3.1. Numerical examples

We examine the performance of the six previously described variance estimators in a variety of settings based upon the bone marrow transplant example of Section 1. In all settings, we simulated times to death from an exponential distribution with mean 10 and censoring times that were uniform over the interval [0, 10]. Times to relapse were simulated from a Weibull distribution with hazard $\lambda_r(t) = \kappa_r\rho_r(\rho_r t)^{\kappa_r - 1}$ and times to aGVHD were simulated from a Weibull distribution with hazard $\lambda_a(t) = \kappa_a\rho_a(\rho_a t)^{\kappa_a - 1}$. The parameters κ_r, ρ_r, κ_a and ρ_a varied in each of four settings so that we could examine the impact of the shape of the corresponding hazards on the variance estimators. The actual parameter values are displayed in Table I.

In settings A and B, relapse had a decreasing hazard, while in settings C and D, relapse had an increasing hazard. Settings A and B differed from each other and settings C and D differed from each other by having aGVHD occur with a decreasing hazard in settings A and C and an increasing hazard in settings B and D. We also examined settings in which either or both

DOI: 10.1002/sim

Table I. Parameter values for each setting.

	Relapse		aGVHD	
Setting	κ_r	ρ_r	κ_a	ρ_a
A	0.5	0.2	0.5	0.2
B	0.5	0.2	2.0	0.2
C	2.0	0.2	0.5	0.2
D	2.0	0.2	2.0	0.2

events had constant hazards, but the results of those settings are not shown as they replicate our findings from settings A–D. We note that in actuality, presence of a GVL effect would induce dependence between the time to aGVHD and the time to relapse. However, as stated earlier, the joint distribution of two dependent risks can be represented by the joint distribution of two independent risks [2].

In each setting, we simulated data for 1000 samples of size $n \in \{20, 50, 100\}$. All estimates for the cumulative incidence of relapse and their variances in each simulated data set were computed at two time points, $t = 1.0$ and 3.0, and the values we report are the averages over all 1000 simulations. We also computed the empirical variance of the 1000 cumulative incidence estimates as a reference point for the variance estimators. For each simulated data set, we also computed a bootstrap-based variance estimate as an alternative to the variance estimators, similar in approach to that of Yuen *et al.* [29]. The number of bootstrap samples in each simulation was set at $B = 200$, a value which has been found to be an adequate number of replications for computing variance estimates [30]. When using Lin's method, we used $B = 500$ resampled values of the data in each simulation, mimicking the value of B used in his manuscript. All computations were done in SAS, with the exception of Gray's variance estimator, which was computed in R using the *cuminc* function contained in the library *comprisk*. A copy of the SAS program containing macros for the remaining variance estimators is available from the authors upon request.

Our results are shown in Tables II and III. In Table II we see the performance of the variance estimators when relapse has a decreasing hazard. In both settings A and B, all the variance estimators tend to perform better at $t = 1$ than at $t = 3$ when $N = 20$ or 50. As relapse has a decreasing hazard, and compounded with the censoring and competing events of aGVHD and death, estimation becomes more difficult at later timepoints as fewer and fewer relapses occur. This finding is less apparent at $N = 100$, which appears to generate a sufficient number of relapses for accurate estimation, even at later timepoints.

In both settings A and B, the bootstrap and the variance estimators of Gaynor *et al.* and Betensky and Schoenfeld perform the best of all the estimators, although they both tend to underestimate the variance in small samples ($N = 20, 50$). The variance estimator of Korn and Dorey does as well as those just mentioned when $t = 1.0$ but slightly underperforms at $t = 3.0$. The variance estimators of Pepe and Lin both substantially underestimate the actual variance for all values of N, with Lin's approach outperforming Pepe's approach. Gray's variance estimator overestimates the actual variance, although the bias is substantially reduced at $N = 100$. In comparing settings A and B, it appears that all the variance estimators perform slightly better with an increasing hazard for aGVHD (Setting B) in small samples ($N = 20$), although this trend is not apparent with larger sample sizes in which none of the variance estimators appear to be affected by the hazard for aGHVD.

DOI: 10.1002/sim

Table II. Comparison of variance estimates for cumulative incidence estimator for event with decreasing hazard. Parameters for both settings defined in Table I. N = number of observations; $\widehat{\mathrm{CI}}(t)$ = estimated cumulative incidence; B = Bootstrap; Gr = Gray; P = Pepe; L = Lin; Ga/B = Gaynor/Betensky; K = Korn; VIF = variance inflation factor proposed for Pepe variance estimator.

				Empirical variance	% of empirical variance						
Setting	N	t	$\widehat{\mathrm{CI}}(t)$		B	Gr	P	L	Ga/B	K	VIF
A	20	1.0	0.2867	0.0106	94.2	105.2	67.1	86.0	94.1	94.3	1.23
		3.0	0.3709	0.0129	93.1	107.6	57.7	80.1	91.8	90.6	1.57
	50	1.0	0.2878	0.0043	97.2	103.9	85.2	93.9	97.4	97.2	1.08
		3.0	0.3711	0.0051	97.5	104.1	79.0	91.8	97.3	95.3	1.25
	100	1.0	0.2897	0.0021	99.6	102.6	92.1	96.3	99.1	98.6	1.04
		3.0	0.3719	0.0026	98.7	101.2	89.6	95.2	98.3	97.0	1.10
B	20	1.0	0.3478	0.0117	95.8	105.5	72.1	88.7	95.3	95.5	1.16
		3.0	0.4775	0.0139	95.9	106.3	63.5	84.3	94.3	92.7	1.46
	50	1.0	0.3460	0.0047	98.7	102.9	88.4	95.5	98.7	98.3	1.08
		3.0	0.4758	0.0057	95.8	105.3	80.7	91.0	95.4	92.8	1.24
	100	1.0	0.3491	0.0024	99.6	100.9	94.0	97.9	99.1	98.7	1.04
		3.0	0.4778	0.0028	100.7	102.5	92.0	97.8	100.4	98.1	1.11

Table III. Comparison of variance estimates for cumulative incidence estimator for event with increasing hazard. Parameters for both settings defined in Table I. N = number of observations; $\widehat{\mathrm{CI}}(t)$ = estimated cumulative incidence; B = Bootstrap; Gr = Gray; P = Pepe; L = Lin; Ga/B = Gaynor/Betensky; K = Korn; VIF = variance inflation factor proposed for Pepe variance estimator.

				Empirical variance	% of empirical variance						
Setting	N	t	$\widehat{\mathrm{CI}}(t)$		B	Gr	P	L	Ga/B	K	VIF
C	20	1.0	0.0277	0.0015	93.8	106.8	65.8	85.6	93.1	92.5	1.16
		3.0	0.1413	0.0078	91.4	107.2	45.2	73.9	88.3	84.8	1.47
	50	1.0	0.0262	0.0006	98.3	105.9	85.5	94.7	98.0	97.6	1.08
		3.0	0.1377	0.0029	99.3	105.5	74.7	92.0	98.3	94.5	1.24
	100	1.0	0.0257	0.0003	100.8	103.9	95.8	99.4	100.8	100.4	1.04
		3.0	0.1388	0.0014	99.3	104.8	88.7	96.2	98.8	96.9	1.11
D	20	1.0	0.0357	0.0018	97.3	103.7	75.3	90.1	96.2	95.6	1.10
		3.0	0.2128	0.0105	95.5	105.1	59.0	82.1	92.8	87.8	1.35
	50	1.0	0.0360	0.0007	98.1	102.4	89.0	95.4	98.1	97.8	1.08
		3.0	0.2136	0.0042	96.0	103.6	80.0	91.0	95.5	92.0	1.25
	100	1.0	0.0356	0.0004	100.4	102.1	95.9	99.0	100.7	100.3	1.04
		3.0	0.2147	0.0021	99.0	101.4	90.4	96.6	99.0	97.7	1.11

DOI: 10.1002/sim

In Table III we see the performance of the variance estimators when relapse has an increasing hazard. As with Table II, we see that the bootstrap and methods of Gaynor *et al.* and Betensky and Schoenfeld outperform the others, closely followed by that of Korn and Dorey, with the method of Gray overestimating the actual variance and the methods of Lin and Pepe underestimating the actual variance. All of the estimators tend to perform better at $t = 1.0$ than at $t = 3.0$, regardless of whether the aGVHD hazard is increasing or decreasing. As most of the findings from Table II are replicated in Table III, it appears that the variance estimators are affected little by the pattern of relapse (increasing or decreasing hazard). However, in setting C, where aGVHD has a decreasing hazard, we see that the overestimation by Gray's variance estimator is further amplified beyond that seen in Setting A (Table II), indicating a possible sensitivity of Gray's method in small samples to the direction of the relapse hazard.

3.2. Proposed inflation factor for Pepe estimator

As seen in our simulations, the method of Pepe tends to significantly underestimate the variance of $\widehat{\mathrm{CI}}(t)$. This finding is not unexpected, as Pepe summarized the results from a small simulation study in her original manuscript and stated that her variance estimator for the KM estimator tended to underestimate the true variance of the KM estimator. However, the degree of the underestimation for the variance of $\widehat{\mathrm{CI}}(t)$ in our simulations was much higher than we anticipated.

Pepe also compared her estimator to Greenwood's formula for the variance of the KM estimator, stating that Greenwood's formula was less biased. Therefore, we postulated that a rough adjustment to Pepe's variance estimator for $\widehat{\mathrm{CI}}(t)$ could be generated from its relation to Greenwood's formula as applied to the KM estimator. Therefore, if we let $V_{\mathrm{P}}(t)$ denote Pepe's variance estimate for the KM estimator at time t and $V_{\mathrm{G}}(t)$ denote Greenwood's variance estimate for the KM estimator at time t, we have a variance-inflation factor (VIF)

$$\begin{aligned}\mathrm{VIF}(t) &= V_{\mathrm{G}}(t)/V_{\mathrm{P}}(t)\\ &= \frac{\int_0^t [Y^2(u) - Y(u)]^{-1}\,\mathrm{d}N(u)}{\int_0^t [Y^{-2}(u) - Y^{-3}(u)]\,\mathrm{d}N(u)}\end{aligned} \tag{7}$$

where $Y(u)$ and $N(u)$ are the martingale quantities defined in Section 2 (see p. 773 of Pepe's manuscript).

We applied this VIF to our simulated data and the average VIFs are shown in the last columns of Tables II and III. In large samples ($N = 100$), we see the proposed VIFs are sufficient to make Pepe's variance estimator consistent with the empirical variance. However, in small samples ($N = 20$), the proposed VIFs are woefully insufficient for making the Pepe estimator consistent with the empirical variance, and in moderately-size samples ($N = 50$), the proposed VIFs greatly improve the Pepe estimator, but are still not quite sufficient to make the Pepe estimator consistent with the empirical variance.

3.3. Actual application

We have a sample of 137 subjects with acute myelocytic leukaemia who participated in a multi-center trial designed to study a conditioning regimen of busulfan and cyclophosphamide before bone marrow transplantation. The specific data can be found in the text of Klein and Moeschberger [31], and details of the study can be found in Copelan *et al.* [32]. The subjects were stratified into three

DOI: 10.1002/sim

Table IV. Comparison of cumulative incidence of GVHD and corresponding standard error estimates in 137 bone marrow transplant patients, 38 with ALL, 54 with low risk AML, and 45 with high risk AML; $\widehat{\mathrm{CI}}(t)$ = estimated cumulative incidence; B = Bootstrap; Gr = Gray; P = Pepe; L = Lin; Ga/B = Gaynor/Betensky; K = Korn.

			Standard error					
t	Disease	$\widehat{\mathrm{CI}}(t)$	B	Gr	P	L	Ga/B	K
100	ALL	0.211	0.068	0.067	0.061	0.068	0.066	0.066
	AML	0.051	0.024	0.022	0.020	0.021	0.022	0.022
	Low risk	0.037	0.027	0.026	0.024	0.026	0.026	0.026
	High risk	0.067	0.041	0.038	0.033	0.037	0.037	0.037
180	ALL	0.395	0.081	0.081	0.069	0.077	0.079	0.080
	AML	0.253	0.046	0.044	0.039	0.042	0.044	0.044
	Low risk	0.222	0.057	0.057	0.051	0.050	0.057	0.057
	High risk	0.289	0.071	0.069	0.055	0.067	0.068	0.068
365	ALL	0.507	0.082	0.085	0.070	0.074	0.081	0.082
	AML	0.384	0.047	0.049	0.045	0.045	0.049	0.049
	Low risk	0.370	0.065	0.067	0.060	0.064	0.066	0.066
	High risk	0.400	0.072	0.075	0.059	0.072	0.073	0.074

groups: 38 patients with acute lymphoblastic leukaemia (ALL), 54 patients with low-risk acute myelogenous leukaemia (AML), and 45 patients with high-risk AML. We stratified our estimates for the cumulative incidence of GVHD and the corresponding standard errors by the two disease groups, as well as by risk group within AML. Thus, our results, as shown in Table IV, are based upon moderate sample sizes and are comparable to the results in Tables II and III with $N = 50$.

Note there is no evidence that these data follow the same distributions as the data used in the simulations. In AML patients (low- and high-risk combined), the risks of relapse and non-GVHD death appear to be relatively flat over the first 365 days, while the risk of GVHD tends to increase over the same time period. In contrast, in ALL patients the risks of relapse and non-GVHD death tend to increase over the first 100 days after transplant and then flatten, while the risk of GVHD tends to continually increase over the first 365 days after relapse.

A more significant difference between this data and that used in the simulations relates to the magnitude of the GVHD hazard to the hazards of the competing events of relapse and non-GVHD death. At the earliest time point examined in this data ($t = 100$), the competing risks hazards tend to be much closer in magnitude to the hazard of GVHD than at the earliest time point ($t = 1$) in the simulated data. Furthermore, the competing risks hazards in this data tend to be less than the hazard of GVHD at $t = 100$, while in the simulated data, the competing risks hazards tend to be greater than the hazard of GVHD at $t = 1$. As a result, this data has a relatively larger number of observed GVHD events at early time points than the simulated data. Such a finding is not unexpected, as GVHD has a much higher frequency than relapse early after transplant, and GVHD is the leading cause of early death in transplant patients.

At $t = 100$, we see that Pepe's method predictively produces lower standard errors than those of the bootstrap and the methods of Gaynor *et al.*, Betensky and Schoenfeld, and Korn and Dorey. However, the standard errors using the methods of Gray and Lin are both surprisingly close to those of Gaynor *et al.*, Betensky and Schoenfeld, and Korn and Dorey. We suspect this finding, which

DOI: 10.1002/sim

differs from the findings from our simulations, is due to the properties of the hazards discussed earlier. It is possible that a large disparity in hazard functions, present in the simulations but not in this data, may impact the performance of Lin's and Gray's variance estimators, although this speculation would need to be confirmed with more research. Nonetheless, at $t = 180$, we start to see indications that Lin's method produces underestimated standard errors, and at $t = 365$, we start to see indications that Gray's method produces inflated standard errors, both of which are supported by our simulation results.

4. CONCLUDING REMARKS

The results of our simulation study and data analysis indicate two important facts: (1) the variance estimator of Pepe is quite biased and leads to underestimated standard errors in small and moderate samples of data, and (2) the estimators of Gaynor *et al.*, Betensky and Schoenfeld, and Korn and Dorey do much better than those of Pepe, Lin, and Gray, and are fairly accurate even in small samples. Although we attempted to improve the performance of the Pepe estimator through an inflation factor, this factor tended to be satisfactory only in large samples and was not adequate in small samples.

It remains unclear why the estimators of Pepe and Lin, although based upon similar theoretic approaches, tend to underestimate the true variance to distinctly different degrees. One explanation is that the approximations used by each author tend to converge at slightly different rates and that Pepe's approximation should be expanded to include higher-order terms. One simple adjustment to Lin's approach would be to sample random values from a central t-distribution rather than a standard normal distribution, although it is unclear what the appropriate degrees of freedom for the t-distribution should be.

It is concerning that Gray's estimator is the only estimator available for general use, as it tends to lead to inflated standard errors. Such overestimation could lead to an increased rate of type II errors (i.e. less power) in studies whose primary outcome is cumulative incidence. We recommend that the variance estimators of Gaynor *et al.*, Betensky and Schoenfeld, and/or Korn and Dorey be included in future statistical packages that include procedures for cumulative incidence estimation.

REFERENCES

1. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 1958; **53**:457–481.
2. Tsiatis A. A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences of the United States of America* 1975; **72**:20–22.
3. Peterson AV. Bounds for a joint distribution function with fixed sub-distribution functions: applications to competing risks. *Proceedings of the National Academy of Sciences* 1976; **73**:11–13.
4. Dignam J, Weissfeld LA, Anderson SJ. Methods for bounding the marginal survival distribution. *Statistics in Medicine* 1995; **14**:1985–1998.
5. Horowitz MM, Gale RP, Sondel PM, Goldman JM, Kersey J, Kolb HJ, Rimm AA, Ringden O, Rozman C, Speck B, Truitt RL, Zwaan FE, Bortin MM. Graft-versus leukaemia reactions following bone marrow transplantation in humans. *Blood* 1989; **75**:555–562.
6. Tai BC, Machin D, White I, Gebski V. Competing risks analysis of patients with osteosarcoma: a comparison of four different approaches. *Statistics in Medicine* 2001; **20**:661–684.
7. Kim W, Poterucha JJ, Benson JT, Therneau TM. The impact of competing risks on the observed rate of chronic hepatitis C progression. *Gastroenterology* 2004; **127**:749–755.

DOI: 10.1002/sim

8. Kaempchen S, Guenther T, Toschke M, Grunkemeier GL, Wottke M, Lange R. Assessing the benefit of biological valve prostheses: cumulative incidence (actual) vs. Kaplan–Meier (actuarial) analysis. *European Journal of Cardio-Thoracic Surgery* 2003; **23**:710–713.
9. Bailey RC, Jia-Yeong Lin M, Krakauer H. Time-to-event modeling of competing risks with intervening states in transplantation. *American Journal of Transplantation* 2003; **3**:192–202.
10. Caplan RJ, Pajak TF, Cox JD. Analysis of the probability and risk of cause-specific failure. *International Journal of Radiation Oncology*, *Biology*, *Physics* 1994; **29**:1183–1186.
11. Benichou J, Gail MH. Estimates of absolute cause-specific risk in cohort studies. *Biometrics* 1990; **46**:379–386.
12. Korn EL, Dorey FJ. Application of crude incidence curves. *Statistics in Medicine* 1992; **11**:813–829.
13. Gaynor JJ, Feuer EJ, Tan CC, Wu DH, Little CR, Straus DJ, Clarkson BD, Brennan MF. On the use of cause-specific failure and conditional failure probabilities: examples from clinical oncology data. *Journal of the American Statistical Association* 1993; **88**:400–409.
14. Bentzen SM, Vaeth M, Pedersen DE, Overgaard J. Why actuarial estimate should be used in reporting late normal-tissue effects of cancer treatment. *International Journal of Radiation Oncology*, *Biology*, *Physics* 1995; **32**:1531–1534.
15. Chappell R. Re: Caplan *et al.* and Bentzen *et al. International Journal of Radiation Oncology*, *Biology*, *Physics* 1996; **36**:988–989.
16. Farley TMM, Mohamed MA, Slaymaker E. Competing approaches to analysis of failure times with competing risks. *Statistics in Medicine* 2001; **20**:3601–3610.
17. Pepe MS, Mori M. Kaplan–Meier, marginal, or conditional probability curves in summarizing competing risk failure time data? *Statistics in Medicine* 1993; **12**:737–751.
18. Aalen O. Nonparametric estimation of partial transition probabilities in multiple decrement models. *Annals of Statistics* 1978; **6**:534–545.
19. Aalen O, Johansen S. An empirical transition matrix for nonhomogeneous Markov chains based on censored observations. *Scandinavian Journal of Statistics* 1978; **5**:141–150.
20. Andersen PK, Borgan O, Gill R, Keiding N. *Statistical Methods Based on Counting Processes*. Springer: Berlin, 1993.
21. Gooley TA, Leisenring W, Crowley J, Storer BE. Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. *Statistics in Medicine* 1999; **18**:695–706.
22. Betensky RA, Schoenfeld DA. Nonparametric estimation in a cure model with random cure times. *Biometrics* 2001; **57**:282–286.
23. Gray RJ. A class of *k*-sample tests for comparing the cumulative incidence of a competing risk. *Annals of Statistics* 1988; **16**:1141–1154.
24. Pepe MS. Inference for events with dependent risks in multiple endpoint studies. *Journal of the American Statistical Association* 1991; **86**:770–778.
25. Lin D. Non-parametric inference for cumulative incidence functions in competing risks studies. *Statistics in Medicine* 1997; **16**:901–910.
26. Greenwood M. The natural duration of cancer. In *Reports on Public Health and Medical Subjects 33*. His Majesty's Stationery Office: London, 1926; 1–26.
27. Dinse GE, Larson MG. A note on semi-Markov models for partially censored data. *Biometrika* 1986; **73**:379–386.
28. Choudhury JB. Non-parametric confidence interval estimation for competing risks analysis: application to contraceptive data. *Statistics in Medicine* 2002; **21**:1129–1144.
29. Yuen KC, Zhu L, Zhang D. Comparing *k* cumulative incidence functions through resampling methods. *Lifetime Data Analysis* 2002; **8**:401–412.
30. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Chapman & Hall: London, 1993.
31. Klein JP, Moeschberger ML. *Survival Analysis*: *Techniques for Censored and Truncated Data* (2nd edn). Springer: Berlin, 2003.
32. Copelan EA, Biggs JC, Thompson JM, Crilley P, Szer J, Klein JP, Kapoor N, Avalos BR, Cunningham I, Atkinson K, Downs K, Harmon GS, Daly MB, Brodsky I, Bulova SI, Utuschka PJ. Treatment for acute myelocytic leukaemia with allogeneic bone marrow transplantation following preparation with Bu/Cy. *Blood* 1991; **78**:838–843.

DOI: 10.1002/sim