BIOMETRICS 56, 173–182
March 2000

Variance and Sample Size Calculations in Quality-of-Life–Adjusted Survival Analysis (Q-TWiST)

Susan Murray

Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.
email: skmurray@umich.edu

and

Bernard Cole

Department of Community and Family Medicine, Dartmouth Medical School,
Lebanon, New Hampshire 03756, U.S.A.

SUMMARY. The Quality-Adjusted Time Without Symptoms or Toxicity (Q-TWiST) statistic previously introduced by Glasziou, Simes and Gelber (1990, *Statistics in Medicine* **9**, 1259–1276) combines toxicity, disease-free survival, and overall survival information in assessing the impact of treatments on the lives of patients. This methodology has received positive reviews from clinicians as intuitive and useful, but to date, the variance of this statistic has remained unspecified. We review aspects of the Q-TWiST method for analyzing clinical trial data, extend the method to accommodate multiple treatment arms, and provide closed-form asymptotic variance formulas. We also provide a framework for designing Q-TWiST clinical trials with sample sizes determined using the derived asymptotic variance formulas. Trials currently collecting quality of life data did not have the benefit of these sample size calculation techniques in designing their studies.

KEY WORDS: Dependent Kaplan–Meier estimates; Quality of life; Sample size; Survival.

1. Introduction

In making decisions about treatment effectiveness in clinical trials, several endpoints may be of interest. Many trials focus on overall survival (OS), but often simplifying the experiences of patients with this single endpoint is less than the whole story. For instance, in breast cancer clinical trials, it is interesting to compare the time a patient lives without recurrence of disease after initial treatment, or disease-free survival (DFS). Clinicians and patients are also interested in the duration of treatment-induced toxicity. The International Breast Cancer Study Group (IBCSG) Trial V collected data on all of these endpoints for the purpose of comparing node-positive breast cancer patients randomized to receive as adjuvant treatments long-duration (6–7 months, $N = 816$) or short-duration (1 month, $N = 413$) chemotherapy (Ludwig Breast Cancer Study Group, 1988; Gelber et al., 1992a). Inherent in collecting information on survival endpoints related to the end of toxicity, to recurrence, and to death is the notion that quality of life (QOL) varies in the stages demarked by these events. A complete picture of a patient's experiences on study cannot be established without considering each of these aspects simultaneously. In the IBCSG trial mentioned above, there was interest in evaluating gains in DFS and OS for the long-duration chemotherapy in light of its additional toxicity. Standard methods available for analyzing multiple endpoints are not designed to detect and resolve treatment differences that occur in different directions. Incorporating knowledge related to QOL in an analysis can be helpful in making treatment recommendations.

One avenue for exploring treatment effects has been proposed by Glasziou, Simes, and Gelber (1990) and is called Q-TWiST, or Quality-Adjusted Time Without Symptoms or Toxicity. The statistic was originally introduced as an extension to the Quality-Adjusted Life Year (QALY) method used in cost-effective analyses with uncensored endpoints, which attributes QOL weights between zero and one to distinct stages of a person's experience on study in producing an adjusted measure of survival time. However, in the nonparametric setting with bounded censoring, the unrestricted mean may not be identifiable. Using the same intuition of weighting the average time spent in each health state according a QOL measure between zero and one, Glasziou et al.'s Q-TWiST statistic allows for asymptotically unbiased estimation of the average QOL-adjusted survival time accumulated by some time, τ, in the presence of censored endpoint data.

In Section 2, we describe and extend methods for Q-TWiST analysis of clinical trial data. To make inferences, suitable variance estimates of the Q-TWiST statistics are required. Currently, the bootstrap method has been employed without evidence that it is correctly estimating the variance. Section 3

presents a closed-form asymptotic variance for the Q-TWiST statistic and recommended estimates. In deriving this asymptotic variance, we also derive the closed-form asymptotic covariance between dependent Kaplan–Meier (KM) estimates or between dependent Nelson–Aalen (NA) hazard estimates. Methods for estimating these dependent KM asymptotic covariances without an exact closed form have been proposed by Wei and Lachin (1984) in relation to their nonparametric tests for equality of multivariate survival endpoints. Our more precise form for these covariances eases mathematical manipulation in Q-TWiST variance calculations required for this research. Additional precursor work characterizing the joint distribution of dependent survival endpoint random variables is well summarized by Anderson et al. (1993). In particular, Prentice and Cai (1992) and Dabrowska (1988) describe nonparametric methods for estimating the joint distribution of two failure time endpoints.

Having a closed-form asymptotic variance improves our ability to make use of the Q-TWiST statistic in an inferential setting. In Section 4, we outline a strategy for determining sample sizes needed for detecting differences in QOL using the Q-TWiST statistic in clinical trials. Using the new Q-TWiST variance, an analysis is performed in Section 5. A discussion follows in Section 6.

2. The Q-TWiST Methodology

The Q-TWiST approach to analyzing clinical trial data is appropriate in many disease settings. However, to simplify the presentation of methodology, we present Q-TWiST in the context of comparing adjuvant therapies for breast cancer, as in IBCSG Trial V. At the beginning of the study, the patients have already undergone surgery to remove all detectable cancer and are subsequently randomized to adjuvant therapies to attack remaining micro–metastatic cancer.

The first step in applying Q-TWiST is to define clinical endpoints that mark changes in a patient's QOL. Define T_1 as the duration of toxicity (TOX), which occurs from the beginning of the study until the end of treatment. Let T_2 be time to disease relapse and T_3 be time to death. Hence $T_2 - T_1$ is the time without symptoms or toxicity (TWiST) that a patient experiences after chemotherapy and $T_3 - T_2$ is the time a person lives after disease recurrence (REL). Together, the three mutually exclusive states, TOX, TWiST, and REL, describe a patient's studytime experience. Although these endpoints are correlated, we assume they are not competing in nature. These methods should not be applied to multiple endpoints where observing one endpoint precludes the observation of another. Note that the methods described here generalize easily to circumstances where fewer or more health states are required to describe the course of disease.

For each treatment group, g, we estimate a QOL-adjusted statistic $Q_g = \mu_{\mathrm{TOX}g} \int_0^\tau \hat{S}_{1g}(t)dt + \int_0^\tau \{\hat{S}_{2g}(t) - \hat{S}_{1g}(t)\}dt + \mu_{\mathrm{REL}g} \int_0^\tau \{\hat{S}_{3g}(t) - \hat{S}_{2g}(t)\}dt$, where $\hat{S}_{ig}(t)$ is the KM estimate for $S_{ig}(t) = P(T_{ig} > t)$, $g = 1, \ldots, G$, $i = 1, \ldots, 3$, and $\mu_{\mathrm{TOX}g}$ and $\mu_{\mathrm{REL}g}$ are weights between zero and one. The upper limit of integration, τ, is chosen so that KM curves are consistent estimates for survival in the area of integration. Glasziou et al. (1990) suggest the median follow-up time as a reasonable choice. Other investigators may choose to maximize the region of area under the survival curve in choosing

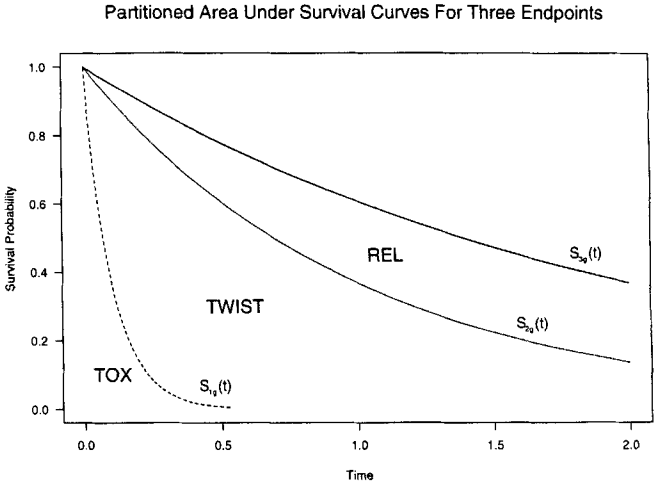

Figure 1. Area between survival curves corresponding to the TOX, TWiST, and REL states.

τ, which would increase power to detect late-term differences in Q-TWiST. For notational convenience, we have restricted the upper limits of Q_g to be the same, but the method extends to situations where the upper limits of integration differ. The statistic, Q_g, reflects the area under the OS curve differentially weighted across mutually exclusive partitions corresponding to TOX, TWiST, and REL as depicted in Figure 1 with $\tau = 2$ years.

The adjustment for QOL on each treatment arm in each survival state is incorporated through the weights $\mu_{\mathrm{TOX}g}$ and $\mu_{\mathrm{REL}g}$. If each weight equals one, Q_g reduces to the unweighted area under the OS curve and the analysis is based on survival time from study entry regardless of QOL. If each weight equals zero, Q_g presents an analysis driven by length of TWiST so that no survival benefit is allowed for time patients spend in toxicity or in relapse. In most cases, weights are chosen to reflect some reduced benefit for life lived under treatment toxicity or relapse.

Various procedures for assigning the weights $\mu_{\mathrm{TOX}g}$ and $\mu_{\mathrm{REL}g}$ are currently in practice or development. Several studies conducted by the IBCSG have begun collecting QOL information longitudinally from patients in order to estimate $\mu_{\mathrm{TOX}g}$ and $\mu_{\mathrm{REL}g}$. However, QOL data is not routinely collected at this time, so many studies that could benefit from investigating QOL issues do not have data available to estimate weights. In this case, a sensitivity analysis may be done displaying results under various weighting scenarios. One advantage of this approach is that treatment recommendations can be tailored to individual QOL perceptions.

Let $n = \Sigma_{g=1}^G n_g$ and $\bar{Q} = (\Sigma_{g=1}^G n_g Q_g)/n$, where n_g is the number of patients on treatment g. Under the null hypothesis (H_0) of no treatment difference in this QOL-adjusted setting, $Q = (Q_1 - \bar{Q}, Q_2 - \bar{Q}, \ldots, Q_{G-1} - \bar{Q})'$ is asymptotically multivariate normal with mean vector $\mathbf{0}$ and variance–covariance Σ. A test statistic for comparing the G treatment groups is $\chi^2 = Q'\Sigma^{-1}Q$, which is asymptotically chi-squared with $G-1$ d.f. The form of Σ will depend on whether fixed weights are used as part of a sensitivity analysis or whether the weights

are estimated from data, in which case modification to accommodate the variation in the weights is required. This will be discussed further in Section 3, where a closed form for Σ is presented for both situations.

It is instructive to look at the special case with no censoring and $\tau = \infty$. Define $Q_{gi} = \mu_{\mathrm{TOX}} T_{1gi} + (T_{2gi} - T_{1gi}) + \mu_{\mathrm{REL}}(T_{3gi} - T_{2gi})$, where $g = 1, \ldots, G$ and $i = 1, \ldots, n_g$. For each individual i on treatment g, Q_{gi} can be interpreted as the length of life a patient lives adjusted for QOL in the different health states. Here $Q_g = (\Sigma_{i=1}^{n_g} Q_{gi})/n_g$ and χ^2 reduces to a multivariate test comparing means of the individual QOL scores Q_{gi} across treatments $g = 1, \ldots, G$. For two treatment groups, the QOL statistic corresponds to the two sample t-test on the Q_{gi}'s, $g = 1, 2$.

3. Derivation of the Variance–Covariance Matrix, Σ, of the Quality-of-Life–Adjusted Q-TWiST Statistic

In deriving a closed form for Σ, a closed-form asymptotic covariance between dependent KM survival estimates will be useful. Suppose we are interested in the covariance of KM estimators for two of the three dependent event times, T_i and T_j, of the last section, with $i, j \in \{1, 2, 3\}$. For each event time T_i, we define a censoring random variable U_i, a failure indicator $\Delta_i = I(T_i < U_i)$, and observable event time $X_i = \min(T_i, U_i)$. We assume $(U_i, i = 1, \ldots, 3)$ are independent of all failure random variables $(T_i, i = 1, \ldots, 3)$. However, there is no restriction on the dependence between U_i and U_j or between T_i and T_j for $i \neq j$. Often U_1, U_2, and U_3 will be exactly equal to one another. However, we allow a more general case where one endpoint, such as toxicity, is censored due to poor record keeping while other endpoint information is available.

In describing hazard functions used in the asymptotic covariances of KM estimates, $\hat{S}_i(x)$ and $\hat{S}_j(y)$, we borrow terminology from Fleming and Harrington (1991) and Anderson et al. (1993). We denote the usual marginal hazards at time u for T_i and T_j with $\lambda_i(u)$ and $\lambda_j(u)$, respectively. Let the crude joint hazard $\lambda_{ij}(u, v) = \lim_{\Delta u, \Delta v \to 0} P(u \leq X_i < u + \Delta u, v \leq X_j < v + \Delta v, \Delta_i = 1, \Delta_j = 1 \mid X_i \geq u, X_j \geq v)/(\Delta u \Delta v)$. A crude conditional hazard function $\lambda_{i|j}(u \mid v) = \lim_{\Delta u \to 0} P(u \leq X_i < u + \Delta u, \Delta_i = 1 \mid X_i \geq u, X_j \geq v)/\Delta u$ associated with time u for T_i has a risk set restricted to patients with $X_i \geq u$ and $X_j \geq v$. Similarly, define $\lambda_{j|i}(v \mid u) = \lim_{\Delta v \to 0} P(v \leq X_j < v + \Delta v, \Delta_j = 1 \mid X_i \geq u, X_j \geq v)/\Delta v$. Since our primary interest in estimation relates to the marginal hazards of T_i and T_j, we do not need to make additional censoring assumptions relating the crude conditional and joint hazards to the usual net conditional and joint hazards for T_i and T_j. For the purposes of describing the variance, accommodating such interpretations are unnecessary.

Define

$$G_{ij}(u, v) = \frac{P(X_i \geq u, X_j \geq v)}{P(X_i \geq u) P(X_j \geq v)} \times \Big[\lambda_{ij}(u, v) - \lambda_{i|j}(u \mid v)\lambda_j(v) - \lambda_{j|i}(v \mid u)\lambda_i(u) + \lambda_i(u)\lambda_j(v)\Big]$$

whenever the risk sets for events i and j are not empty at times u and v, respectively, or $G_{ij}(u, v) = 0$ otherwise. It is shown in Appendix A that

$$\begin{aligned}\sigma_{S_{ij}}(t_1, t_2) &= \operatorname{cov}\left[\sqrt{n}\left\{\hat{S}_i(t_1) - S_i(t_1)\right\}, \sqrt{n}\left\{\hat{S}_j(t_2) - S_j(t_2)\right\}\right] \\ &= S_i(t_1) S_j(t_2) \int_0^{t_1} \int_0^{t_2} G_{ij}(u, v) dv du. \qquad (1)\end{aligned}$$

An estimate for this covariance is also provided in Appendix A. As an additional check on our derivation, we have verified several familiar special cases. For instance, when X_i and X_j are independent, this covariance becomes zero. Also, when X_i is identically equal to X_j, this covariance reduces to the variance of the KM estimate. With uncensored data, where each KM estimate reduces to a simple proportion, (1) becomes $P(T_i > t_i, T_j > t_j) - P(T_i > t_i)P(T_j > t_j)$, which corresponds to the covariance for two dependent proportions. With this asymptotic covariance in closed form, we proceed to derive the covariance, Σ, from Section 2.

Adding a subscript, g, denoting treatment group to previous notation, additional calculations in Appendix A show that the covariance of restricted means estimates, $\int_0^\tau \hat{S}_{ig}(t_1) \times dt_1$ and $\int_0^\tau \hat{S}_{jg}(t_2) dt_2$, is V_{ijg}/n_g, where

$$V_{ijg} = \int_0^\tau \int_0^\tau \left[\int_v^\tau S_{jg}(t) dt\right] \left[\int_u^\tau S_{ig}(t) dt\right] G_{ijg}(u, v) dv du.$$

In terms of restricted means, Q_g may be rewritten as

$$\begin{aligned}&\left(\mu_{\mathrm{TOX}g} - 1\right) \int_0^\tau \hat{S}_{1g}(t) dt \\ &\quad + \left(1 - \mu_{\mathrm{REL}g}\right) \int_0^\tau \hat{S}_{2g}(t) dt + \mu_{\mathrm{REL}g} \int_0^\tau \hat{S}_{3g}(t) dt.\end{aligned}$$

Hence, when the weights $\mu_{\mathrm{TOX}g}$ and $\mu_{\mathrm{REL}g}$ are considered fixed, as in a sensitivity analysis,

$$\operatorname{var}(Q_g) = \sum_{i=1}^{3} \sum_{j=1}^{3} w_{ig} w_{jg} V_{ijg}/n_g = V_g/n_g,$$

where $w_{1g} = \mu_{\mathrm{TOX}g} - 1$, $w_{2g} = 1 - \mu_{\mathrm{REL}g}$, $w_{3g} = \mu_{\mathrm{REL}g}$, and $V_g = \Sigma_{i=1}^3 \Sigma_{j=1}^3 w_{ig} w_{jg} V_{ijg}$. From this, we find that the ℓ, mth element of Σ from Section 2 in the fixed weights case is

$$\sigma_{\ell m} = n_\ell^{-1} V_\ell \left[I(\ell = m) - n^{-1} n_\ell\right] - n^{-1} V_m + n^{-2} \sum_{g=1}^{G} n_g V_g,$$

where $I(\ell = m)$ is equal to one when $\ell = m$ and is equal to zero otherwise.

The form of Σ needs to be modified when $\mu_{\mathrm{REL}g}$ and $\mu_{\mathrm{TOX}g}$ are estimated from data. For each treatment group g, let $\hat{w}_{ig}$ estimate w_{ig}, where $\mathrm{E}(\hat{w}_{ig}) = w_{ig}$. We denote $\operatorname{cov}(\hat{w}_{ig}, \hat{w}_{jg})$ by $\sigma_{w_{ijg}}/n_g$ and $\hat{\boldsymbol{w}}_{\boldsymbol{g}} = (\hat{w}_{1g}, \hat{w}_{2g}, \hat{w}_{3g})'$. We leave $\hat{\boldsymbol{w}}_{\boldsymbol{g}}$ vaguely specified since we do not wish to restrict the user from using any of the variety of methods potentially employed to obtain these estimates. For example, if little QOL data is recorded on each individual, then a population averaged estimate stratified by treatment group and health state may be of interest. If multiple QOL measurements are collected on individuals within each health state, then a more subject-specific estimate of these weights might be desirable. Of course, different estimation procedures for $\hat{\boldsymbol{w}}_{\boldsymbol{g}}$ would result in different

estimates of $\sigma_{w_{ijg}}/n_g$. An assumption is made that estimated weights are independent of time spent in health states.

In this case, $Q_g = \Sigma_{i=1}^{3} \hat{w}_{ig} \int_0^\tau \hat{S}_{ig}(t)dt$ and

$$\begin{aligned}\text{var}(Q_g) &= \text{var}(\text{E}(Q_g \mid \hat{\boldsymbol{w}}_{\boldsymbol{g}})) + \text{E}(\text{var}(Q_g \mid \hat{\boldsymbol{w}}_{\boldsymbol{g}})) \\ &= \text{var}\left(\sum_{i=1}^{3} \hat{w}_{ig} \int_0^\tau S_{ig}(t)dt\right) \\ &\quad + \text{E}\left(\sum_{i=1}^{3}\sum_{j=1}^{3} \hat{w}_{ig}\hat{w}_{jg}V_{ijg}/n_g\right) \\ &= \sum_{i=1}^{3}\sum_{j=1}^{3}\left\{\int_0^\tau S_{ig}(t)dt\right\}\left\{\int_0^\tau S_{jg}(t)dt\right\} \\ &\quad \times \sigma_{w_{ijg}}/n_g \\ &\quad + \sum_{i=1}^{3}\sum_{j=1}^{3}\{(\sigma_{w_{ijg}}/n_g) + w_{ig}w_{jg}\}V_{ijg}/n_g.\end{aligned}$$

As expected, this result reduces to the case where weights are known when $\sigma_{w_{ijg}}$ approaches zero. In cases where the weights are determined from a population-based average estimate, $\sigma_{w_{ijg}}/n_g^2$ is relatively small and the above can be simplified to

$$\begin{aligned}&\sum_{i=1}^{3}\sum_{j=1}^{3}\left\{\int_0^\tau S_{ig}(t)dt\right\}\left\{\int_0^\tau S_{jg}(t)dt\right\}\sigma_{w_{ijg}}/n_g + V_g/n_g \\ &\quad = V_g^*/n_g,\end{aligned}$$

where

$$\begin{aligned}V_g^* = \sum_{i=1}^{3}\sum_{j=1}^{3}\Bigg[&\sigma_{w_{ijg}}\left\{\int_0^\tau S_{ig}(t)dt\right\}\left\{\int_0^\tau S_{jg}(t)dt\right\} \\ &+ (w_{ig}w_{jg})V_{ijg}\Bigg].\end{aligned}$$

So the ℓ, mth element of Σ from Section 2 in this case is

$$\begin{aligned}\sigma^*_{\ell m} &= n_\ell^{-1}V_\ell^*\left[I(\ell = m) - n^{-1}n_\ell\right] \\ &\quad - n^{-1}V_m^* + n^{-2}\sum_{g=1}^{G} n_g V_g^*.\end{aligned}$$

Recommended estimates for all variances from this section are located in Appendix B.

4. Study Design Considerations and Simulation Results

This section provides an example of how to calculate sample sizes for detecting differences in quality of life using the Q-TWiST statistic for two treatment group comparisons. We will discuss both the case where fixed weights are to be used and the case where population-based weights are estimated from patient data collected during the trial. In either case, the investigator should have input as to the number of health states patients experience, the projected survival behavior of endpoints defining these states, and the restriction time, τ.

Let n be the desired sample size in each of two treatment groups being compared. Borrowing selected notation from previous sections, the test statistic using fixed weights may be represented as $Z = n^{1/2}(Q_1 - Q_2)/(V_1 + V_2)^{1/2}$, which asymptotically behaves as a standard normal random variable. Note that if weights are estimated, we replace V_1 and V_2 with V_1^* and V_2^* to account for extra variability. Under the alternative hypothesis, $Q_1 - Q_2$ has a nonzero mean, Δ, which measures the true difference in QOL-adjusted survival. Hence, Z has asymptotic mean $n^{1/2}\Delta/(V_1 + V_2)^{1/2}$. The distance between this alternative mean, $n^{1/2}\Delta/(V_1 + V_2)^{1/2}$, and the zero mean, which holds true under H_0, can also be represented by $z_{\alpha/2} + z_\beta$, where α is the type I error chosen for the test, β is the type II error, and z_* represents the percentile of the standard normal distribution that cuts off area $*$ in the upper tail. From this observation and some further algebra, we see that, for detecting the alternative Δ with power $1-\beta$ and size α, the asymptotic relationship $n = \{(z_{\alpha/2}+z_\beta)/\Delta\}^2(V_1+V_2)$ must be true or, specifically, in the fixed weights case,

$$\begin{aligned}n = &\left(\frac{z_{\alpha/2} + z_\beta}{\Delta}\right)^2 \\ &\times \sum_{g=1}^{2}\sum_{i=1}^{3}\sum_{j=1}^{3} w_{ig}w_{jg}\int_0^\tau\int_0^\tau\left[\int_v^\tau S_{jg}(t)dt\right] \\ &\qquad \times \left[\int_u^\tau S_{ig}(t)dt\right] \\ &\qquad \times G_{ijg}(u,v)dvdu, \qquad (2)\end{aligned}$$

and in the estimated population-based weights case,

$$\begin{aligned}n = &\left(\frac{z_{\alpha/2} + z_\beta}{\Delta}\right)^2 \\ &\times \sum_{g=1}^{2}\sum_{i=1}^{3}\sum_{j=1}^{3}\Bigg[\sigma_{w_{ijg}}\left\{\int_0^\tau S_{ig}(t)dt\right\} \\ &\qquad \times \left\{\int_0^\tau S_{jg}(t)dt\right\} + (w_{ig}w_{jg})V_{ijg}\Bigg]. \qquad (3)\end{aligned}$$

One may proceed determining sample sizes using these asymptotic formulas.

To demonstrate the use of these sample size formulas, we plan a QOL study where patients experience three different QOL states marked by survival endpoints, T_{1g}, T_{2g}, and T_{3g} for the two treatment groups $g = 1, 2$, where $T_{1g} \leq T_{2g} \leq T_{3g}$. In each treatment group, the toxicity duration, T_{1g}, is taken to be independent of the length of time spent in the other two health states and is distributed as uniform$(0, 1/6)$. Hence, the simulated duration of toxicity corresponds to a duration incurred during a 2-month treatment course within each group.

Two sources of correlation between T_{2g} and T_{3g} are incorporated into this simulation. The first source of correlation comes from the ordered nature of the survival endpoints, so larger observed T_{2g} tend to result in larger observed values of T_{3g}. The second source of correlation comes from the relationship between T_{2g}^* and T_{3g}^*, the durations of time within the second and third health state, respectively, for those in group g. For instance, one might believe that a patient with large T_{2g}^* will also tend to have larger T_{3g}^*. To model the second source of correlation, we assume the distribution of the length of

Table 1
Estimated sample sizes for various censoring levels and treatment differences

Alternative μ's	Censoring percentage[a]	$\mu_{TOX} = 0.5$ $\mu_{REL} = 0.5$	$\mu_{TOX} = 1.0$ $\mu_{REL} = 1.0$	$\mu_{TOX} = 0$ $\mu_{REL} = 0$	$E(\mu_{TOX}) = 0.5$ $E(\mu_{REL}) = 0.5$
(−0.20, −0.20)	0	420	474	440	421
	20	448	520	474	454
	40	484	581	518	485
(−0.25, −0.25)	0	268	301	281	268
	20	286	330	303	286
	40	309	369	332	309
(−0.30, −0.30)	0	185	208	195	186
	20	198	227	210	200
	40	214	254	230	214

[a] Censoring percentage measures the degree of censoring in the study up to time τ among each cohort of patients.

the time in these two health states for each treatment group, T^*_{2g} and T^*_{3g}, to be bivariate log-normal with centrality parameters (μ_{1g}, μ_{2g}), dispersion parameters $(\sigma_{1g}, \sigma_{2g})$, and correlation parameter ρ_g. In each case considered, we choose $(\mu_{11}, \mu_{21}) = (0, 0)$, $(\sigma_{1g}, \sigma_{2g}) = (1, 1), g = 1, 2$ and $\rho = 0.9$. Various (μ_{12}, μ_{22}) parameters are explored in order to study different levels of Q-TWiST treatment differences. Using these various parameters, the survival endpoints for each treatment group g become T_{1g}, $T_{2g} = T_{1g} + T^*_{2g}$ and $T_{3g} = T_{2g} + T^*_{3g}$.

To model the censoring mechanism acting on the endpoints T_{1g}, T_{2g}, and T_{3g}, we defined the random variable $C = U \times I(B = 1) + \tau \times I(B = 0)$, where U is a uniform random variable on $[1/6, \tau]$ and B is a Bernoulli random variable with success probability related to the desired level of censoring. Note that the percentage of censoring reported in Table 1 measures the degree of censoring in the study up to time τ among each cohort of patients.

All of the previously described survival and censoring distributions are used in simulation to estimate V_1, V_2, and Δ for $\tau = 2$ years in (2). Future programmers may decide to automate this procedure somewhat by querying the end user in relation to the marginal distributions for the endpoints and displaying results for a range of correlation parameters and censoring percentages. Since true values of V_1 and V_2 are not sample size-dependent, a sample size of 250 per treatment group is used to obtain estimates of V_1 and V_2 as described in Section 3 and then averaged over 5000 simulations of the data. To estimate Δ, algebraic simplifications for $\hat{\Delta}$ in the fully observed data case can be exploited since Δ is not affected by censoring levels. In this case, for any KM estimator $\hat{S}(t)$ on failure times t_j with n_τ failures occurring prior to time τ, $\int_0^\tau \hat{S}(t)dt = \Sigma_{t_j < \tau} t_j/n + \tau\{1 - (n_\tau/n)\}$. For this simple calculation, it is possible to use a large sample size of 250,000 to estimate Δ. In order to make $\hat{\Delta}$ even more precise, a Monte Carlo simulated average of $\hat{\Delta}$ over 2500 simulations with this sample size was calculated. Using this method, $\hat{\Delta}$ was accurate to roughly four decimal places. In Table 1, estimated Δ's for $(\mu_{12}, \mu_{22}) = (-0.20, -0.20)$ with $(\mu_{TOX}, \mu_{REL}) = \{(0.5, 0.5), (1.0, 1.0), (0.0, 0.0)\}$ were (0.11148, 0.10331, 0.11965). Estimated Δ's for these weights with $(\mu_{12}, \mu_{22}) = (-0.25, -0.25)$ were (0.13997, 0.13068, 0.14926) and with $(\mu_{12}, \mu_{22}) = (-0.30, -0.30)$ were (0.16867, 0.15861, 0.17872).

Table 1 displays sample sizes required for 80% power and 5% type I error as determined from (2) or (3), as appropriate, under various censoring levels, (μ_{12}, μ_{22}) parameters, and weighting choices. Note that the sample sizes displayed assume a single statistical test. If multiple tests are used in a sensitivity analysis, the type I error may be adjusted according to the user's favorite method. Although this is an interesting and important topic, we will not devote any time discussing the issue of multiple comparisons, which is well documented elsewhere (e.g., Bickel and Doksum, 1977). The four weighting choices displayed give a flavor for the different possible analyses using Q-TWiST. The first set of weights assigns $\mu_{TOX} = \mu_{REL} = 0.5$, which penalizes the TOX and REL states for reduced QOL in these time intervals. The second set of weights assigns $\mu_{TOX} = \mu_{REL} = 1$, which reduces to a two-sample test comparing the restricted means for OS. The third displayed set of weights assigns $\mu_{TOX} = \mu_{REL} = 0$ so that the analysis is driven by the time without symptoms or toxicity, completely discounting time spent in the TOX and REL health states in the analysis. The last column of sample sizes in Table 1 is based on the case where population-based weights are estimated from patient data. These estimated weights were simulated as normal random variables with mean 0.5 and variance $0.05/n$ in each treatment group, resulting in identical $\hat{\Delta}$'s to the fixed weight case with $\mu_{TOX} = \mu_{REL} = 0.5$ and largely similar estimated sample sizes.

Figure 2 displays the survival endpoints for the smallest Q-TWiST treatment difference studied. Notice that the partitioned survival area corresponding to the TWiST health state is more favorable for the group with survival endpoints displayed on the left. Also, the area in the REL health state appears to be only slightly different in the two treatment groups.

One thousand Monte Carlo simulations were conducted to verify 80% power and 5% type I error planned for the analysis for selected entries of Table 1. For instance, using sample sizes (268, 286, 309) under $(\mu_{12}, \mu_{22}) = (-0.25, -0.25)$ with censoring percentages (0%, 20%, 40%), simulated power was

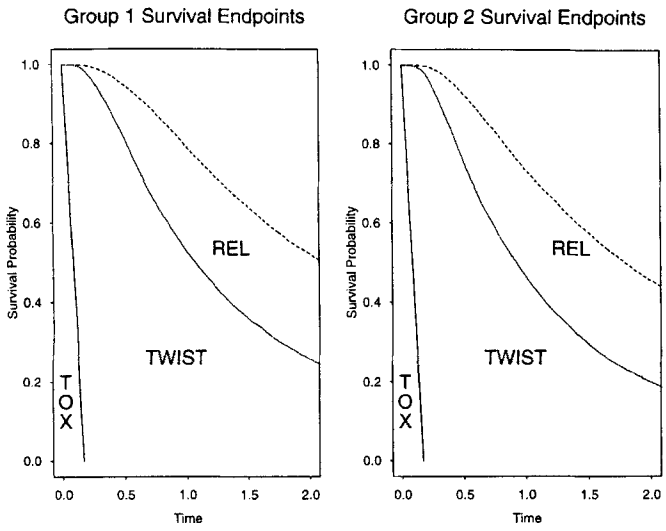

Figure 2. Area between simulated survival curves corresponding to the TOX, TWiST, and REL states for $\mu_{\text{TOX}} = \mu_{\text{REL}} = 0.5$ and $\Delta = 0.11148$.

(79.4%, 77.7%, 80.5%) with corresponding type I error under H_0 of (4.9%, 5.5%, 5.6%) in the fixed weights case and (79.6%, 77.5%, 80.3%) with corresponding type I error under H_0 of (4.7%, 5.7%, 5.4%) in the estimated population-based weights case. Empirical Q-TWiST variance estimates corresponding to these simulations matched closely with the simulation-specific Q-TWiST variance estimates. As an example, with 0% censoring with $(\mu_{12}, \mu_{22}) = (-0.25, -0.25)$, the empirical variance across all simulations was 0.00245, whereas the (25%, 50%, 75%) quantiles for the observed Q-TWiST estimated variances using these formulas were (0.00243, 0.00249, 0.00256). Under H_0, the empirical variance of 0.00248 compares favorably with the observed quantiles (0.00233, 0.00240, 0.00246). Results displayed in Table 1 and in further unreported simulations demonstrate that incorporating survival information related to the TOX and REL states through partial weighting of these states in the Q-TWiST statistic increases the chances of detecting the simulated treatment differences. Interestingly, the statistic based on time spent in the TWiST state alone is almost as efficient as the statistic giving partial weights to the remaining states in this scenario. This happens primarily because the difference between partitioned areas under the survival curve for the two treatments occurs mainly within the TWiST health state for these selected distributions.

5. Example

We now return to the IBCSG Trial V breast cancer study comparing long- versus short-duration chemotherapy mentioned in the introduction. Three health states were identified for the analysis: (1) time with toxicity due to chemotherapy (TOX), (2) time without toxicity or disease relapse (TWiST), and (3) time following disease relapse (REL).

For the short-duration chemotherapy group, restricted mean estimates of TOX, DFS, and OS within the 84-month median follow-up period were observed to be 0.85, 48.13, and 63.97 months with variances 0.00127, 2.35330, and 1.52404. Covariance estimates of these restricted means were

$$(\hat{V}_{121}/413, \hat{V}_{131}/413, \hat{V}_{231}/413) = (0.00281, 0.00137, 1.46990).$$

In the long-duration group, TOX, DFS, and OS 84-month restricted means were 5.79, 59.30, and 68.52 months with variances 0.00932, 1.04318, and 0.71836 and covariances

$$(\hat{V}_{122}/816, \hat{V}_{132}/816, \hat{V}_{232}/816) = (0.00722, 0.00254, 0.74252).$$

Hence, the long-duration chemotherapy regimen has a longer duration of toxicity as well as increased DFS and OS. A sensitivity analysis considering a large variety of weights provides the best perspective on overall treatment benefit when no QOL data is available for estimating the weights and also allows a clinical practitioner to assess potential treatment benefit profiles for patients with different perceived QOL. For a patient with high QOL in all disease stages, the usual analysis using $\mu_{\text{TOX}} = \mu_{\text{REL}} = 1$ gives an average of 4.55 months of life gained on the long-duration chemotherapy in relation to the short-duration therapy during the first 84 months on study (95% CI: 1.616, 7.484). For a patient with little tolerance for toxicity, using $\mu_{\text{TOX}} = \mu_{\text{REL}} = 0$, we find an average of 6.23 months of quality-adjusted life gained on the long-duration chemotherapy in these 84 months (95% CI: 2.624, 9.836).

In Table 2, we display mean estimated quality-adjusted survival differences between the long- and short-duration chemotherapies in months along with standard errors for a variety of weights considered by Gelber et al. (1991, 1995). We also include bootstrap-based standard errors as used in their original analyses. In their papers, the weights μ_{TOX} and μ_{REL} were assumed to be equivalent across treatment groups. Inferences using either the bootstrap or asymptotic closed-form variance estimates are similar. For most scenarios, the long-duration chemotherapy provided a significantly longer QOL-adjusted mean survival, even when adjusted for multiple comparisons. Cases where the sensitivity analysis did not distinguish a treatment preference involved weights that highly penalized toxicity while simultaneously judging near perfect QOL for the relapse health state. In the limited patient preference data available on these types of patients, these weight choices would not be typical. In fact, most patients rate the REL health state as inferior to the TOX health state in terms of QOL in these breast cancer studies. Hence, an analysis of this type would support assigning most patients to the long-duration chemotherapy.

In this example, the bootstrap method provides estimates similar to the closed-form-based estimates. Currently there is no research available to justify the appropriateness of bootstrapping covariances based on dependent marginal KM-based estimators such as we use in this research. In other work related to multiple failure-time endpoints, Yandell and Horvath (1988) demonstrated that the covariance of bivariate survival estimators can be successfully bootstrapped in cases where the true covariance is complex. This may turn out to be the case with dependent marginal survival estimates as well.

6. Discussion

The evaluation of treatments in terms of QOL is becoming increasingly important in clinical research. In an article in the *Journal of Clinical Oncology*, O'Shaughnessy et al. (1991),

Table 2
Q-TWiST sensitivity analysis for various weights

μ_{TOX}	μ_{REL}	Q-TWiST difference[a]	SE (variance formula)	SE (bootstrap)[b]	ΔSE
1.00	1.00	4.55	(1.497)	(1.513)	−0.016
1.00	0.75	6.21	(1.518)	(1.519)	−0.001
1.00	0.50	7.86	(1.586)	(1.575)	0.011
1.00	0.25	9.52	(1.697)	(1.675)	0.022
1.00	0.00	11.17	(1.843)	(1.813)	0.030
0.75	1.00	3.32	(1.497)	(1.513)	−0.016
0.75	0.75	4.97	(1.517)	(1.519)	−0.002
0.75	0.50	6.63	(1.585)	(1.574)	0.011
0.75	0.25	8.28	(1.696)	(1.675)	0.021
0.75	0.00	9.94	(1.842)	(1.812)	0.030
0.50	1.00	2.08	(1.497)	(1.513)	−0.016
0.50	0.75	3.74	(1.517)	(1.519)	−0.002
0.50	0.50	5.39	(1.585)	(1.574)	0.011
0.50	0.25	7.05	(1.695)	(1.674)	0.021
0.50	0.00	8.70	(1.841)	(1.812)	0.029
0.25	1.00	1.50	(1.497)	(1.514)	−0.017
0.25	0.75	2.50	(1.517)	(1.519)	−0.002
0.25	0.50	4.16	(1.585)	(1.574)	0.011
0.25	0.25	5.81	(1.695)	(1.674)	0.021
0.25	0.00	7.47	(1.840)	(1.812)	0.028
0.00	1.00	−0.39	(1.498)	(1.515)	−0.017
0.00	0.75	1.27	(1.518)	(1.520)	−0.002
0.00	0.50	2.92	(1.585)	(1.575)	0.010
0.00	0.25	4.58	(1.695)	(1.675)	0.020
0.00	0.00	6.23	(1.840)	(1.812)	0.028

[a] Long duration minus short duration (in months).
[b] The bootstrap method used 1000 iterations in estimating variances.

who hold research positions in the Food and Drug Administration or the National Cancer Institute, pressed for statistical analyses of drug performance that incorporate multiple endpoints such as DFS, OS, and QOL in determining treatment recommendations. Already, the Q-TWiST technique has become a popular tool for this purpose. The basic methodology has been applied in a number of analyses of clinical trials. Gelber, Goldhirsch, and Cavalli (1991) and Gelber et al. (1992a) present analyses of adjuvant therapies for operable breast cancer, and Gelber et al. (1992b) and Lenderking et al. (1994) present analyses of zidovudine therapy for HIV infection. Cole et al. (1996) evaluated the risks and benefits of high-dose interferon alfa-2b adjuvant treatment for malignant melanoma. Gelber et al. (1996) evaluated chemotherapy plus radiation therapy for rectal cancer. In addition, a number of methodological extensions to Q-TWiST have been proposed. Incorporation of covariates and prognostic factors in a Q-TWiST analysis have been proposed by Cole, Gelber, and Goldhirsch (1993) using proportional hazards models and by Cole, Gelber, and Anderson (1994) using accelerated failure time models. An overview of recent extensions is given by Gelber et al. (1995).

In a related recent extension of the original Q-TWiST methodology, Zhao and Tsiatis (1997) discuss a consistent estimator for the distribution, as opposed to the mean, of a lifetime adjusted for known QOL weights along with a consistent variance estimate without requiring a progression of health states as in the original paper by Glasziou, Simes, and Gelber (1990) or this work. Their distribution relates to the distribution of the Q-TWiST statistic in the case where the known weights are piecewise constant. An advantage to the variance calculations based on partitioned survival as in this work is that this approach can easily accommodate estimated weights, while the estimator of Zhao and Tsiatis remains valid only with known weights. Since estimated weights are critical when studying patients' perceived QOL, this is an important distinction.

This research expands the methodological framework of the Q-TWiST procedures by providing closed-form variances of the treatment-specific QOL-adjusted survival estimates and test statistics. Simulations using the newly derived variance show that, in addition to a more complete assessment of a treatment's performance, we may in fact increase our power to detect clinical differences when QOL considerations are included in an analysis. In addition, this research outlines an example of how to design a clinical trial with appropriate sample sizes to detect clinical differences using the Q-TWiST method. Currently the clinical trials collecting QOL information have not had the benefit of these strategies for assuring an adequately powered study.

Acknowledgements

The authors thank R. Strawderman, A. A. Tsiatis, and R. Gelber for helpful conversations and also thank the IBCSG for the use of their data. This work was supported by grant

PBR-53 and RPG-90-013-08-PBP from the American Cancer Society and grant CA-75362 from the National Cancer Institute.

Résumé

La statistique Q-TWIST, précédemment introduite par Glasziou, Simes et Gelber (1990, *Statistics in Medicine* **9,** 1259–1276), combine toxicité, survie sans maladie et information globale sur la survie dans l'évaluation de l'impact des traitements sur la vie des patients. Cette méthodologie a été considérée intuitive et utile par les cliniciens, mais la variance de cette statistique n'a toujours pas été spécifiée. Nous commentons quelques aspects de la méthode Q-TWIST dans l'analyse des données issues d'essais thérapeutiques, nous étendons cette méthode à plusieurs bras de traitement, et nous donnons la variance asymptotique sous forme approchée. Nous donnons également un cadre pour définir des essais thérapeutiques adaptés au Q-TWIST avec des tailles d'échantillon obtenues en utilisant les formules obtenues pour la variance asymptotique. La définition des études recueillant des données de qualité de vie ne bénéficie pas de l'avantage procuré par ces méthodes de calcul de taille d'échantillon.

References

Anderson, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* New York: Springer-Verlag.

Bickel, P. J. and Doksum, K. A. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics.* San Francisco: Holden-Day.

Breslow, N. and Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship. *Annals of Statistics* **2,** 437–453.

Cole, B. F., Gelber, R. D., and Goldhirsch, A. for the International Breast Cancer Study Group (IBCSG). (1993). Cox regression models for quality-adjusted survival analysis. *Statistics in Medicine* **12,** 975–987.

Cole, B. F., Gelber, R. D., and Anderson, K. M. for the International Breast Cancer Study Group (IBCSG). (1994). Parametric approaches to quality-adjusted survival analysis. *Biometrics* **50,** 621–631.

Cole, B. F., Gelber, R. D., Kirkwood, J. M., Goldhirsch, A., Barylak, E., and Borden, E. (1996). A quality-of-life–adjusted survival analysis of interferon alfa-2b adjuvant treatment for high-risk resected cutaneous melanoma: An ECOG study (E1684). *Journal of Clinical Oncology* **14,** 2666–2673.

Dabrowska, D. M. (1988). Kaplan–Meier estimate on the plane. *Annals of Statistics* **16,** 1475–1489.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis.* New York: Wiley.

Gelber, R. D., Goldhirsch, A., and Cavalli, F. for the International Breast Cancer Study Group (IBCSG). (1991). Quality-of-life–adjusted evaluation of a randomized trial comparing adjuvant therapies for operable breast cancer. *Annals of Internal Medicine* **114,** 621–628.

Gelber, R. D., Goldhirsch, A., Hürny, C., Bernhard, J., and Simes, R. J. for the International Breast Cancer Study Group (IBCSG). (1992a). Quality of life in clinical trials of adjuvant therapies. *Journal of the National Cancer Institute Monographs* **11,** 127–135.

Gelber, R. D., Lenderking, W. R., Cotton, D. J., Cole, B. F., Fischl, M. A., Goldhirsch, A., and Testa, M. A. for the AIDS Clinical Trials Group. (1992b). Quality-of-life evaluation in a clinical trial of zidovudine therapy in patients with mildly symptomatic HIV infection. *Annals of Internal Medicine* **116,** 961–966.

Gelber, R. D., Cole, B. F., Gelber, S., and Goldhirsch, A. (1995). Comparing treatments using quality-adjusted survival: The Q-TWiST method. *American Statistician* **49,** 161–169.

Gelber, R. D., Goldhirsch, A., Cole, B. F., Wieand, H. S., Schroeder, G., and Krook, J. E. (1996). A quality-adjusted time without symptoms or toxicity (Q-TWiST) analysis of adjuvant radiation therapy and chemotherapy for resectable rectal cancer. *Journal of the National Cancer Institute* **88,** 1039–1045.

Glasziou, P. P., Simes, R. J., and Gelber, R. D. (1990). Quality adjusted survival analysis. *Statistics in Medicine* **9,** 1259–1276.

Lenderking, W. R., Gelber, R. D., Cotton, D. J., Cole, B. F., Goldhirsch, A., Volderding, P. A., and Testa, M. A. (1994). Evaluation of the quality-of-life assessment in asymptomatic human immunodeficiency virus infection. *New England Journal of Medicine* **330,** 738–743.

Ludwig Breast Cancer Study Group. (1988). Combination adjuvant chemotherapy for node-positive breast cancer: Inadequacy of a single perioperative cycle. *New England Journal of Medicine* **319,** 677–683.

O'Shaughnessy, J. A., Wittes, R. E., Burke, G., Friedman, M. A., Johnson, J. R., Niederhuber, J. E., Rothenberg, M. L., Woodcock, J., Chabner, B. A., and Temple, R. (1991). Commentary concerning demonstration of safety and efficacy of investigational anticancer agents in clinical trials. *Journal of Clinical Oncology* **9,** 2225–2232.

Prentice, R. L. and Cai, J. (1992). Covariance and survivor function estimation using censored multivariate failure time data. *Biometrika* **79,** 495–512.

Wei, L. J. and Lachin, J. M. (1984). Two-sample asymptotically distribution-free tests for incomplete multivariate observations. *Journal of the American Statistical Association* **79,** 653–661.

Yandell, B. S. and Horvath, L. (1988). Bootstrapped multidimensional product limit process. *Australian Journal of Statistics* **30,** 342–358.

Zhao, H. and Tsiatis, A. (1997). A consistent estimator for the distribution of quality adjusted survival time. *Biometrika* **84,** 339–348.

Received September 1998. *Revised June* 1999.
Accepted July 1999.

Appendix A

Asymptotic Multivariate Distribution of Correlated KM Estimates and Corresponding Restricted Means

Denoting the NA estimate by $\hat{\Lambda}(x)$ and the KM by $\hat{S}(x)$, Breslow and Crowley (1974) show that

$$n^{1/2}\left[\hat{S}(x) - \exp\{-\hat{\Lambda}(x)\}\right] \xrightarrow{\mathcal{P}} 0.$$

Hence, determining the asymptotic behavior of dependent NA estimates will provide us with an understanding of the asymptotic behavior for dependent KM estimates. For each of the dependent event times of interest, $i = 1, 2$, define U_{ik} as the censoring random variable corresponding to T_{ik}, $\Delta_{ik} = I(T_{ik} < U_{ik})$, and $X_{ik} = \min(T_{ik}, U_{ik})$, $k = 1 \cdots n$. Also define $N_i(t) = \Sigma_{k=1}^n I(X_{ik} \leq t, \Delta_{ik} = 1)$ and $Y_i(t) = \Sigma_{k=1}^n I(X_{ik} \geq t)$. Let $M_i(t) = N_i(t) - \int_0^t \lambda_i(u) Y_i(u) du$ be the martingale defined with respect to the filtration containing all available censoring and survival data for the endpoint corresponding to i prior to time t and define $M_j(t) = N_j(t) - \int_0^t \lambda_j(u) Y_j(u) du$ similarly for endpoint type j. The filtrations concerning $M_i(t)$ and $M_j(t)$ are dependent but not necessarily nested. Hence, we explicitly derive covariances relating to these martingales without using the usual strategy of conditioning on a common filtration.

We define $J_i(u) = I(Y_i(u) > 0) = I\{(Y_i(u)/n) > 0\}$ to avoid representing infinite integrands below. Asymptotically, $J_i(u) \xrightarrow{\mathcal{P}} p_{J_i}(u) = I\{P(X_i \geq u) \geq 0\}$ does not contribute additional variability to the statistics of interest and is used for notational convenience. Also note that

$$\begin{aligned} M_i(t) &= N_i(t) - \int_0^t \lambda_i(u) Y_i(u) du \\ &= \sum_{k=1}^n \left[I(X_{ik} \leq t, \Delta_{ik} = 1) - \int_0^t \lambda_i(u) I(X_{ik} \geq u) du \right] \\ &= \sum_{k=1}^n M_{ik}(t) \end{aligned}$$

is a sum of independent and identically distributed quantities.

In terms of the NA estimator, we need to find the covariance of terms taking the form

$$\begin{aligned} \sqrt{n}\left\{\hat{\Lambda}_i(t) - \Lambda_i(t)\right\} &\approx \sqrt{n} \int_0^t J_i(u) \frac{dM_i(u)}{Y_i(u)} \\ &= \sqrt{n} \sum_{k=1}^n \int_0^t J_i(u) \frac{dM_{ik}(u)}{Y_i(u)} \\ &= \sqrt{n} \frac{\sum_{k=1}^n \int_0^t J_i(u) \frac{dM_{ik}(u)}{Y_i(u)/n}}{n}. \end{aligned}$$

Since, according to the Glivenko and Cantelli theorem,

$$\sup_{u \in [0,t)} \left| \frac{Y_i(u)}{n} - P(X_i \geq u) \right| \xrightarrow{\mathcal{P}} 0,$$

we rewrite the above as

$$\sqrt{n}\left(\sum_{k=1}^n \int_0^t J_i(u) \Big[\{Y_i(u)/n\}^{-1} + \{P(X_i \geq u)\}^{-1} - \{P(X_i \geq u)\}^{-1} \Big] dM_{ik}(u)/n \right),$$

which after an application of the martingale central limit theorem (or Lenglart's Inequality) has the same limiting distribution as

$$\sqrt{n}\left[\sum_{k=1}^n \int_0^t p_{J_i}(u) \{P(X_i \geq u)\}^{-1} dM_{ik}(u)/n \right].$$

The multivariate central limit theorem identifies the covariance of interest as

$$\operatorname{cov}\left[\int_0^{t_1} p_{J_i}(u) \{P(X_i \geq u)\}^{-1} dM_{ik}(u), \int_0^{t_2} p_{J_j}(v) \left\{P(X_j \geq v)\right\}^{-1} dM_{jk}(v) \right].$$

Since each of the terms above has expectation zero, this covariance may be written as

$$\begin{aligned} \mathrm{E}\Bigg(&\left[\int_0^{t_1} p_{J_i}(u) \{P(X_i \geq u)\}^{-1} dM_{ik}(u) \right] \\ &\times \left[\int_0^{t_2} p_{J_j}(v) \left\{P(X_j \geq v)\right\}^{-1} dM_{jk}(v) \right] \Bigg) \\ &= \int_0^{t_1} \int_0^{t_2} A_{ij}(u, v) \left\{P(X_i \geq u, X_j \geq v)\right\}^{-1} \\ &\qquad \times \mathrm{E}\left\{dM_{ik}(u) dM_{jk}(v)\right\}, \end{aligned}$$

where

$$A_{ij}(u, v) = p_{J_i}(u) p_{J_j}(v) \frac{P(X_i \geq u, X_j \geq v)}{P(X_i \geq u) P(X_j \geq v)}.$$

Note that

$$\begin{aligned} dM_{ik}(u) \approx &\lim_{\Delta u \to 0} I(u \leq X_{ik} < u + \Delta u, \Delta_{ik} = 1) \frac{du}{\Delta u} \\ &- \lambda_i(u) I(X_{ik} \geq u) du, \end{aligned}$$

where $du/\Delta u = 1 + o(du)$. Similarly,

$$\begin{aligned} dM_{jk}(v) \approx &\lim_{\Delta v \to 0} I(v \leq X_{jk} < v + \Delta v, \Delta_{jk} = 1) \frac{dv}{\Delta v} \\ &- \lambda_j(v) I(X_{jk} \geq v) dv. \end{aligned}$$

Substituting these expressions for $dM_{ik}(u)$ and $dM_{jk}(v)$ and taking the expectation through the limit via the dominated convergence theorem, where $R_{i,j}(u, v) = P(X_i \geq u, X_j \geq v)$,

$$\begin{aligned} &\sigma_{\Lambda_{ij}}(t_1, t_2) \\ &= \operatorname{cov}\left[\sqrt{n}\left\{\hat{\Lambda}_i(t_1) - \Lambda_i(t_1)\right\}, \sqrt{n}\left\{\hat{\Lambda}_j(t_2) - \Lambda_j(t_2)\right\} \right] \\ &= \int_0^{t_1} \int_0^{t_2} A_{ij}(u, v) \\ &\qquad \times \lim_{\Delta u, \Delta v \to 0} \Bigg[\frac{B_{i,j}(u, v)}{(\Delta u \Delta v) R_{i,j}(u, v)} \\ &\qquad\quad - \lambda_j(v) \frac{C_{i,j}(u, v)}{(\Delta u) R_{i,j}(u, v)} \\ &\qquad\quad - \lambda_i(u) \frac{D_{i,j}(u, v)}{(\Delta v) R_{i,j}(u, v)} \\ &\qquad\quad + \lambda_i(u) \lambda_j(v) \frac{E_{i,j}(u, v)}{R_{i,j}(u, v)} \Bigg] dv du \end{aligned}$$

$$= \int_0^{t_1}\int_0^{t_2} A_{ij}(u,v) \times \{\lambda_{ij}(u,v) - \lambda_{i|j}(u \mid v)\lambda_j(v) - \lambda_{j|i}(v \mid u)\lambda_i(u) + \lambda_i(u)\lambda_j(v)\}dvdu$$

$$= \int_0^{t_1}\int_0^{t_2} G_{ij}(u,v)dvdu,$$

where

$$B_{i,j}(u,v) = \mathrm{E}\{I(u \le X_{ik} < u + \Delta u, v \le X_{jk} < v + \Delta v, \Delta_{ik} = 1, \Delta_{jk} = 1)\}$$
$$C_{i,j}(u,v) = \mathrm{E}\left\{I(u \le X_{ik} < u + \Delta u, X_{jk} \ge v, \Delta_{ik} = 1)\right\}$$
$$D_{i,j}(u,v) = \mathrm{E}\left\{I(v \le X_{jk} < v + \Delta v, X_{ik} \ge u, \Delta_{jk} = 1)\right\}$$
$$E_{i,j}(u,v) = \mathrm{E}\left\{I(X_{ik} \ge u, X_{jk} \ge v)\right\}.$$

An application of the delta method gives equation (1). This result also leads to the covariance between corresponding restricted mean estimates since

$$\mathrm{cov}\left(\int_0^\tau \hat{S}_i(t_1)dt_1, \int_0^\tau \hat{S}_j(t_2)dt_2\right) = \frac{1}{n}\int_0^\tau\int_0^\tau S_i(t_1)S_j(t_2)\int_0^{t_1}\int_0^{t_2} G_{ij}(u,v)dvdudt_1dt_2$$
$$= \frac{1}{n}\int_0^\tau\int_0^\tau \left[\int_v^\tau S_j(t)dt\right]\left[\int_u^\tau S_i(t)dt\right] G_{ij}(u,v)dvdu$$
$$= \frac{1}{n}V_{ij}.$$

Appendix B

Estimation of Variances

All asymptotic closed-form covariance terms in this manuscript are easily estimated. However, we require additional notation. Let $Y_{ij}(u,v) = \Sigma_{k=1}^n I(X_{ik} \ge u, X_{jk} \ge v)$ count individuals still at risk for both events T_i and T_j. Also, with some abuse of notation, let $dN_{ij}(u,v) = \Sigma_{k=1}^n I(u \le X_{ik} < u + \Delta u, v \le X_{jk} < v + \Delta v, \Delta_{ik} = 1, \Delta_{jk} = 1)$ count individuals with event T_i at time u and event T_j at time v, $dN_{i|j}(u \mid v) = \Sigma_{k=1}^n I(u \le X_{ik} < u + \Delta u, X_{jk} \ge v, \Delta_{ik} = 1)$ count individuals with event T_i at time u who are still at risk for event T_j at time v, and $dN_{j|i}(v \mid u) = \Sigma_{k=1}^n I(v \le X_{jk} < v + \Delta v, X_{ik} \ge u, \Delta_{jk} = 1)$ count individuals with event T_j at time v who are still at risk for event T_i at time u. A simple estimate for $G_{ij}(u,v)$ is

$$\hat{G}_{ij}(u,v) = n\frac{Y_{ij}(u,v)}{Y_i(u)Y_j(v)} \times \left[\frac{dN_{ij}(u,v)}{Y_{ij}(u,v)} - \frac{dN_{i|j}(u \mid v)dN_j(v)}{Y_{ij}(u,v)Y_j(v)} - \frac{dN_{j|i}(v \mid u)dN_i(u)}{Y_{ij}(u,v)Y_i(u)} + \frac{dN_i(u)dN_j(v)}{Y_i(u)Y_j(v)}\right].$$

So

$$\hat{\sigma}_{\Lambda_{ij}}(t_1,t_2) = \sum_{\{u:X_i \le t_1, \Delta_i = 1\}}\sum_{\{v:X_j \le t_2, \Delta_j = 1\}} \hat{G}_{ij}(u,v)$$

and

$$\hat{\sigma}_{S_{ij}}(t_1,t_2) = \hat{S}_i(t_1)\hat{S}_j(t_2)\hat{\sigma}_{\Lambda_{ij}}(t_1,t_2).$$

Adding the subscript g for group,

$$\hat{V}_{ijg} = \sum_{\{u:X_{i(g)} \le \tau, \Delta_{i(g)} = 1\}}\sum_{\{v:X_{j(g)} \le \tau, \Delta_{jg} = 1\}} \left\{\int_v^\tau \hat{S}_{jg}(t)dt\right\} \times \left\{\int_u^\tau \hat{S}_{ig}(t)dt\right\} \times \hat{G}_{ijg}(u,v),$$

where subscripts under the summation signs index the observed failure times for the ith and jth survival endpoints occurring up to and including time τ. Hence,

$$\hat{V}_g = \sum_{i=1}^3\sum_{j=1}^3 w_{ig}w_{jg}\hat{V}_{ijg}$$

and

$$\hat{V}_g^* = \sum_{i=1}^3\sum_{j=1}^3 \left[\hat{\sigma}_{w_{ijg}}\left\{\int_0^\tau \hat{S}_{ig}(t)dt\right\}\left\{\int_0^\tau \hat{S}_{jg}(t)dt\right\} + (\hat{w}_{ig}\hat{w}_{jg})\hat{V}_{ijg}\right]$$

may be used to estimate the elements of Σ as described in Section 3.