Biostatistics (2007), **8**, 2, *pp.* 184–196
doi:10.1093/biostatistics/kxj040
Advance Access publication on April 24, 2006

Parametric regression on cumulative incidence function

JONG-HYEON JEONG*

Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15261, USA
jeong@nsabp.pitt.edu

JASON P. FINE

*Department of Statistics and Department of Biostatistics and Medical Informatics,
University of Wisconsin, Madison, WI 53706, USA*

Summary

We propose parametric regression analysis of cumulative incidence function with competing risks data. A simple form of Gompertz distribution is used for the improper baseline subdistribution of the event of interest. Maximum likelihood inferences on regression parameters and associated cumulative incidence function are developed for parametric models, including a flexible generalized odds rate model. Estimation of the long-term proportion of patients with cause-specific events is straightforward in the parametric setting. Simple goodness-of-fit tests are discussed for evaluating a fixed odds rate assumption. The parametric regression methods are compared with an existing semiparametric regression analysis on a breast cancer data set where the cumulative incidence of recurrence is of interest. The results demonstrate that the likelihood-based parametric analyses for the cumulative incidence function are a practically useful alternative to the semiparametric analyses.

Keywords: Breast cancer; Clinical trial; Competing risks; Cumulative incidence; Cure model; Improper distribution; Regression; Transformation model.

1. Introduction

Competing risks data are encountered frequently in medical research. In breast cancer trials, like Protocol B-19 (Fisher *and others*, 1989) at the National Surgical Adjuvant Breast and Bowel Project (NSABP), a patient may experience multiple events, such as local or regional recurrence, distant metastasis, second primary cancer other than breast, and death. Investigators may be interested in the time and the type of the first event, leading to a competing risks structure. For example, radiation oncologists (Taghian *and others*, 2004) may focus on local or regional recurrences alone as a first event to evaluate radiation therapy after surgery. The cumulative incidence function (Kalbfleisch and Prentice, 1980) quantifies the cumulative probability of cause-specific failure in the presence of competing events without assumptions about the dependence among the events (Korn and Dorey, 1992; Pepe and Mori, 1993; Gaynor *and others*, 1993).

When analyzing cause-specific failure patterns, investigators may be interested in the effects of covariates on the event-specific failure probabilities. Such analyses may involve testing the effects of treatment

*To whom correspondence should be addressed.

adjusted for important prognostic factors, in addition to testing the effects of the prognostic factors. For patient management, it is helpful to characterize the proportion of events over time for causes of interest conditionally on covariates, including treatment. Gray (1988) proposed a nonparametric inference procedure to compare the cumulative incidence curves among two groups, with alternative tests discussed in Pepe (1991). Fine and Gray (1999) and Fine (2001) adapted the proportional hazards model (Cox, 1972, 1975) to the cumulative incidence function and proposed inferences for the effects of treatment and other continuous prognostic factors. In breast cancer trials at NSABP, one may wish to quantify the effect of radiation therapy adjusted for known risk factors, such as age, tumor size, number of positive lymph nodes, and estrogen receptor level. Regression models are critical when adjusting for such variables, which are usually collected as either count or continuous measurements.

Unlike with traditional survival endpoints, the proportion of recurrences in breast cancer tends to increase for a period of time and then plateau (Karrison *and others*, 1999). Those patients who do not experience recurrences can be viewed as a cured population (Boag, 1949) and inferences about the recurrence “distribution” can be conceptualized in a cure model framework. The asymptote of the cumulative distribution function of recurrences is less than 1, that is, the “distribution” is improper. A key issue in regression modeling of recurrences is the effect of covariates on the leveling off point of the cumulative distribution function.

The most widely used analyses of competing risks data in practical applications, like the breast cancer data, are nonparametric and semiparametric. A major advantage of these approaches is that there is no need to assume an underlying distributional form for the cumulative incidence function, which is difficult in the competing risks setting, owing to the impropriety of this function. Of course, such flexibility arises at the cost of efficiency loss relative to parametric models, especially with small sample sizes (Miller, 1983; Jeong and Oakes, 2003). The trade-off for additional parametric assumptions is the potential bias associated with model misspecification (Meier *and others*, 2004). On the other hand, the parametric models permit extrapolation of long-term event probabilities, which are of inherent interest and which cannot generally be identified from nonparametric and semiparametric models. Moreover, parametric regression models are amenable to formal maximum likelihood (and Bayesian) inferences, unlike the semiparametric analyses of Fine and Gray (1999) and Fine (2001).

Jeong and Fine (2006) proposed a direct parameterization of the cumulative incidence function without covariates. Empirical studies using NSABP breast cancer data showed that a simple form of the improper Gompertz distribution (Gompertz, 1825) provided better fits than more complex parametric mixture models (Larson and Dinse, 1985), as measured by the agreement of the fits with nonparametric estimates of the cumulative incidence functions. The direct parameterization is more parsimonious and has a more straightforward interpretation than do indirect parameterizations, where the cause-specific hazard function (Bryant and Dignam, 2004) and/or the overall survival function (Benichou and Gail, 1990) are modeled with proper distributions. This is particularly true if long-term event probabilities are of interest, which cannot be inferred directly from the component models in the indirect parameterization.

In this paper, we extend Jeong and Fine (2006) to the regression setting. Maximum likelihood inferences are developed in which parametric models for the cumulative incidence functions for all causes are fit simultaneously. Inferences are based on standard asymptotic results for the maximum likelihood estimators. Our general parametric framework encompasses models which are parametric specializations of the models in Fine and Gray (1999) and Fine (2001). In the analysis of the Protocol B-19 data set, we employ generalized odds rate regression models (Dabrowska and Doksum, 1998), with Gompertz baseline distributions, including both proportional hazards and proportional odds models as special cases. The proposed parametric procedure permits goodness-of-fit tests for the proportional hazards and proportional odds assumptions, assuming that the parametric model for the base distribution is correctly specified.

In Section 2, we introduce the competing risks data and the associated notation. In Section 3, general parametric regression models are formulated for the cumulative incidence function. The Gompertz distribution is presented and its suitability as a model for the baseline distribution is discussed. In Section 4, maximum likelihood estimation is presented. In Section 5, the variances of the estimated regression parameters and cumulative incidence probabilities are derived. In Section 6, the parametric procedure is compared with the semiparametric Fine–Gray (1999) model on NSABP Protocol B-19.

2. CUMULATIVE INCIDENCE FUNCTION—ONE SAMPLE CASE

The basic identifiable quantities from competing risk data (T, K) are the cause-specific hazard and cumulative incidence functions, where T is time to the first event and $K \in (1, \ldots, n_K)$ is the event type, where n_K is the number of event types. The cause-specific hazard function for an event $K = k$ at time t is $\lambda_k(t) = \lim_{\Delta \to 0} \Pr(t < T \leqslant t + \Delta, K = k | T \geqslant t)/\Delta$. For small Δ,

$$\frac{\Pr(t < T \leqslant t + \Delta, K = k)}{\Delta} \approx \Pr(T \geqslant t)\lambda_k(t). \tag{2.1}$$

As noted in Jeong and Fine (2006), the left-hand side in (2.1) approximates the probability density function for the kth cause-specific event as Δ approaches 0. This implies that the cumulative incidence function for the kth cause-specific event is

$$F_k(t) = \int_0^t S(u)\mathrm{d}\Lambda_k(u), \tag{2.2}$$

where $S(t) = \Pr(T > t)$ and $\Lambda_k(t) = \int_0^t \lambda_k(u)\mathrm{d}u$ is the cumulative hazard function for the kth cause-specific event. In (2.1), the cause-specific hazard function $\lambda_k(t)$ on the right-hand side makes the probability density function for cause-specific events of type k improper whenever $\lambda_k < \sum_k \lambda_k$. Therefore, the cumulative incidence function in (2.2) may also be improper.

In practice, T is typically subject to additional independent right censoring. To nonparametrically estimate the cumulative incidence function, the overall survival function $S(\cdot)$ may be replaced by the Kaplan–Meier estimator (Kaplan and Meier, 1958) and the cause-specific cumulative hazard function $\Lambda_k(\cdot)$ may be replaced by a Nelson–Aalen estimator (Nelson, 1972; Aalen, 1978). Bryant and Dignam (2004) recently proposed a semiparametric inference procedure by parameterizing only the cause-specific hazard function in the integral (2.2), with $S(\cdot)$ being estimated nonparametrically. They noticed an efficiency gain over the nonparametric estimator of F_k. Benichou and Gail (1990) considered fully parametric inference on the cumulative incidence function by parameterizing the cause-specific hazard function and the overall survival function. Larson and Dinse (1985) considered parametric inference for a mixture model representation of the joint distribution of (T, K). Jeong and Fine (2006) proposed direct parametric modeling of $F_k(\cdot)$, and suggested the Gompertz distribution, which is tailored to the unique features of $F_k(\cdot)$. They showed that the direct parametric approach may provide a better fit than either the cause-specific hazard or the mixture model approaches when there is a plateau in the tail of the cumulative incidence function.

3. REGRESSION MODELS

For direct regression modeling of the cumulative incidence function, it is convenient to consider a transformation model structure (Fine and Gray, 1999; Fine, 2001). For events of type k,

$$g_k\{F_k(t; \mathbf{Z})\} = u_k(t) + \mathbf{Z}^T \boldsymbol{\beta}_k, \quad k = 1, \ldots, n_K, \tag{3.1}$$

where $u_k(t)$ is an invertible and monotonically increasing function, $\boldsymbol{\beta}_k$ is a $P \times 1$ parameter vector, and $\mathbf{Z}$ is a time-independent $P \times 1$ covariate vector. For two individuals with covariate vectors $\mathbf{Z}_1$ and $\mathbf{Z}_2$, the conditional cumulative incidence functions satisfy a vertical shift model

$$g_k\{F_k(t; \mathbf{Z}_2)\} - g_k\{F_k(t; \mathbf{Z}_1)\} = (\mathbf{Z}_2 - \mathbf{Z}_1)^T \boldsymbol{\beta}_k. \tag{3.2}$$

Specifying $g_k(\cdot)$ to be the logit function gives a proportional odds model for cause k events, with the regression parameters in $\boldsymbol{\beta}_k$ corresponding to time-independent log odds ratios per unit increases in the covariates.

Fine and Gray (1999) considered the proportional hazards model to directly infer the effects of covariates on the cumulative incidence of type k events. Their model was originally posited in terms of the subdistribution hazard function (Gray, 1988) specifying that

$$\lambda_k^{\mathrm{CI}}(t; \mathbf{Z}) = \lambda_{k0}(t) \exp(\mathbf{Z}^T \boldsymbol{\beta}_k), \tag{3.3}$$

where $\lambda_k^{\mathrm{CI}}(t; \mathbf{Z}) = \lim_{\Delta \to 0} \Pr\{t \leqslant T \leqslant t + \Delta, K = k | T \geqslant t \cup (T \leqslant t \cap K \neq k); \mathbf{Z}\}/\Delta$. The subdistribution hazard function $\lambda_k^{\mathrm{CI}}(\cdot)$ is the hazard function for the improper random variable $T^* = I(K = k) \times T + \{1 - I(K = k)\} \times \infty$, where $I(\cdot)$ is an indicator function. The model (3.3) corresponds to the transformation model (3.1) where $g_k(v) = \log\{-\log(1 - v)\}$. Equivalently, the cumulative probability of a type k event is given by

$$F_k(t; \mathbf{Z}) = 1 - \exp\{-\exp(\mathbf{Z}^T \boldsymbol{\beta}_k) u_k(t)\}, \tag{3.4}$$

where $u_k(t) = \log_k \left\{\int_0^t \lambda_{k0}(s)\mathrm{d}s\right\}$. For semiparametric inference about $\boldsymbol{\beta}_k$ separately from $\lambda_{k0}(t)$ and separately from the models for $F_j(t; \mathbf{Z})$ $(j \neq k)$, Fine and Gray (1999) constructed a partial likelihood in which the risk set for type k events is constructed so that subjects having already experienced events other than type k are always at future "risk" of a type k event. This differs from the traditional cause-specific hazard analysis where the occurrence of an event other than type k removes an individual from future risk sets (Kalbfleisch and Prentice, 1980).

Extending Fine (2001), we propose a general parametric class of transformation models, in which there may be unknown parameters in $g_k(\cdot)$. Each event type has its own model, with distinct parameters for $g_k(\cdot)$, $u_k(\cdot)$, and $\boldsymbol{\beta}_k$. The link function $g_k(\cdot)$ in (3.1) may have arbitrary parametric form $g_k(v_k; \alpha_k)$, where α_k may be unknown. In the Protocol B-19 analysis, we employ the odds rate transformation

$$g_k(v_k; \alpha_k) = \log[\{(1 - v_k)^{-\alpha_k} - 1\}/\alpha_k], \quad -\infty < \alpha_k < \infty, \tag{3.5}$$

which includes the proportional hazards and proportional odds models (Dabrowska and Doksum, 1998) when $\alpha \to 0$ and $\alpha = 1$, respectively. Adopting a flexible link function is useful for assessing the goodness-of-fit of the proportional hazards model and other fully specified models for $g_k(\cdot)$. Under model (3.5), the cumulative probability of a type k event is given by

$$F_k(t; \mathbf{Z}) = 1 - \{1 + \alpha_k \exp(\mathbf{Z}^T \boldsymbol{\beta}_k) u_k(t)\}^{-1/\alpha_k}. \tag{3.6}$$

In this paper, the Gompertz (1825) distribution is used to parameterize the log baseline cumulative subdistribution hazard function, $u_k(t)$, which permits the asymptote of the cumulative distribution function to be <1. The cumulative distribution function can be written

$$B(t; \rho, \tau) = 1 - \exp[\tau\{1 - \exp(\rho t)\}/\rho], \tag{3.7}$$

where $-\infty < \rho < \infty$ and $0 < \tau < \infty$. The hazard function $\{\mathrm{d}B(t; \rho, \tau)/\mathrm{d}t\}\{1 - B(t; \rho, \tau)\}^{-1}$ is

$$\lambda^G(t; \rho, \tau) = \tau \exp(\rho t), \tag{3.8}$$

and hence the cumulative hazard function is given by $u^G(t; \rho, \tau) = \tau\{\exp(\rho t) - 1\}/\rho$. An improper distribution occurs when $\rho < 0$ and $\tau < \infty$. The hazard function (3.8) can fit either increasing or decreasing hazards, depending on the signs of the parameters.

The implied parametric regression model for the cumulative incidence function is

$$F_k(t; \boldsymbol{\psi}_k, \mathbf{Z}) = 1 - \{1 + \alpha_k \exp(\mathbf{Z}^T \boldsymbol{\beta}_k) u_k^G(t; \rho_k, \tau_k)\}^{-1/\alpha_k}, \tag{3.9}$$

where $\boldsymbol{\psi}_k = (\alpha_k, \boldsymbol{\beta}_k^T, \rho_k, \tau_k)$. For large t, one minus the cumulative distribution function in (3.9) equals the proportion of patients cured, given a set of covariate values. With competing risks data, this cured fraction is the proportion of individuals never experiencing event k, for example, breast cancer recurrence. Note that, when $\rho < 0$, as $t \to \infty$, the formula reduces to

$$F_k(\infty; \boldsymbol{\psi}_k, \mathbf{Z}) = 1 - \{1 - \alpha_k \tau_k \exp(\mathbf{Z}^T \boldsymbol{\beta}_k)/\rho_k\}^{-1/\alpha_k}, \tag{3.10}$$

so that the proportion never experiencing a type k event is $1 - F_k(\infty; \boldsymbol{\psi}_k, \mathbf{Z})$.

4. MAXIMUM LIKELIHOOD ESTIMATION

Let T_i and C_i be the potential failure time and the potential censoring time, respectively, for the ith subject. Define $X_i = \min(T_i, C_i)$. The indicators for the competing events are

$$\delta_{ki} = \begin{cases} 1, & \text{if the } i\text{th subject experiences the } k\text{th cause-specific event as a first event,} \\ 0, & \text{otherwise,} \end{cases}$$

for $k = 1, \ldots, n_K$. In the breast cancer example where $n_K = 2$, $\delta_{1i} = 1$ if the ith patient experiences a recurrence as a first event and 0 otherwise, and δ_{2i} is similarly defined for the ith patient experiencing events other than recurrences. The observable data are denoted as $\left(X_i, \delta_{1i}, \ldots, \delta_{n_K i}, \mathbf{Z}_i\right)$ $(i = 1, \ldots, n)$.

Following similar arguments for direct inference for F_k in the one-sample case (Jeong and Fine, 2006), given covariate $\mathbf{Z}_i = \mathbf{z}_i$, the likelihood function is

$$\prod_{i=1}^{n} \left[\left\{ \prod_{k=1}^{n_K} f_k(x_i, \boldsymbol{\psi}_k; \mathbf{z}_i)^{\delta_{ki}} \right\} \left\{ 1 - \sum_{k=1}^{n_K} F_k(x_i, \boldsymbol{\psi}_k; \mathbf{z}_i) \right\}^{1 - \sum_{k=1}^{n_K} \delta_{ki}} \right], \tag{4.1}$$

where $f_k(x, \boldsymbol{\psi}_k; \mathbf{z}_i) = \mathrm{d}F_k(x, \boldsymbol{\psi}_k; \mathbf{z}_i)/\mathrm{d}x$ $(k = 1, \ldots, n_K)$. Note that the likelihood involves information from all failure types and does not factor into separate pieces for each type. This differs from the parameterization based on the cause-specific hazard functions (Prentice *and others*, 1978), where the likelihood factors so that inferences about cause 1 may be carried out separately from the models for other causes. Under this formulation, misspecification of cause-specific hazard models for other causes does not lead to bias in the estimated model for cause 1. A limitation is that direct inference about the cumulative incidence functions is not possible.

In (4.1), the cumulative probability of failure from any event is the sum of the n_K cumulative incidence functions, $F_1(\cdot), \ldots, F_{n_K}(\cdot)$. Under the cause-specific hazard formulation, this cumulative probability is obtained from the overall hazard rate, which is the sum of the corresponding n_K cause-specific hazard functions (Benichou and Gail, 1990; Bryant and Dignam, 2004).

When the proportional subdistribution hazard model (Fine and Gray, 1999) is assumed for type k events, $F_k(x, \boldsymbol{\psi}_k; \mathbf{z}_i)$ may be replaced with $F_k(x; \mathbf{z})$ in (3.4) after parameterizing $u_k(x)$ by $u_k^G(x; \rho_k, \tau_k)$. In this paper, our focus is a model with general $g_k(v_k; \alpha_k)$, including the odds rate model (3.9), which accommodates a range of nonproportional hazards models.

From (4.1), the log-likelihood function is given by

$$\sum_{i=1}^{n} \left[\sum_{k=1}^{n_K} \delta_{ki} \log_k\{ f_k(x_i, \boldsymbol{\psi}_k; \mathbf{z}_i)\} + \left(1 - \sum_{k=1}^{n_K} \delta_{ki}\right) \log \left\{ 1 - \sum_{k=1}^{n_K} F_k(x_i, \boldsymbol{\psi}_k; \mathbf{z}_i) \right\} \right]. \tag{4.2}$$

Differentiating (4.2) and setting the resulting score function equal to 0 with respect to $\boldsymbol{\psi}_k$, the maximum likelihood estimator $\widehat{\boldsymbol{\psi}}_k$, $k = 1, \ldots, n_K$, can be obtained. The maximum likelihood estimator of $F_k(t; \boldsymbol{\psi}_k, \mathbf{z})$ is $F_k(t, \widehat{\boldsymbol{\psi}}_k; \mathbf{z})$, $k = 1, \ldots, n_K$.

5. Large sample inferences

The observed information matrix can be derived by taking the second derivatives of the log-likelihood function. Given $\mathbf{Z} = \mathbf{z}$, the variance of $F_k(t, \widehat{\boldsymbol{\psi}}_k; \mathbf{z})$ can be evaluated by the multivariate delta method as

$$\widehat{\mathrm{var}}\{F_k(t, \widehat{\boldsymbol{\psi}}_k; \mathbf{z})\} = \left(\frac{\partial F_k(t, \boldsymbol{\psi}_k; \mathbf{z})}{\partial \boldsymbol{\psi}_k}\right)\Bigg|_{\boldsymbol{\psi}_k = \widehat{\boldsymbol{\psi}}_k} \widehat{\mathrm{var}}(\widehat{\boldsymbol{\psi}}_k) \left(\frac{\partial F_k(t, \boldsymbol{\psi}_k; \mathbf{z})}{\partial \boldsymbol{\psi}_k}\right)'\Bigg|_{\boldsymbol{\psi}_k = \widehat{\boldsymbol{\psi}}_k}, \tag{5.1}$$

where $\partial F_k(t, \boldsymbol{\psi}_k; \mathbf{z})/\partial \boldsymbol{\psi}_k$ is a vector of the first derivatives of the cumulative incidence function for the kth cause-specific event with respect to $\boldsymbol{\psi}_k$. The matrix $\widehat{\mathrm{var}}(\widehat{\boldsymbol{\psi}}_k)$ is a submatrix of the inverse of the observed information matrix corresponding to the variance of $\widehat{\boldsymbol{\psi}}_k$, evaluated at $\widehat{\boldsymbol{\psi}}_1, \ldots, \widehat{\boldsymbol{\psi}}_{n_K}$. A pointwise 95% confidence interval for $F_k(t; \mathbf{z})$ is

$$F_k(t, \widehat{\boldsymbol{\psi}}_k; \mathbf{z}) \pm 1.96 \times \sqrt{\widehat{\mathrm{var}}\{F_k(t, \widehat{\boldsymbol{\psi}}_k; \mathbf{z})\}}, \quad k = 1, \ldots, n_K. \tag{5.2}$$

Confidence intervals may also be based on inverting the likelihood ratio test, which may have better small sample properties than (5.2). A disadvantage is that these intervals cannot be calculated using output from the fitted model and additional computations are needed.

One may evaluate the regression parameters and the form of $g_k(\cdot)$ with simple Wald-type test statistics. Under model (3.9), the hypotheses are (i) H_0: $\alpha_k = \alpha_{k0}$, $k = 1, \ldots, n_K$, for testing the proportional hazards ($\alpha_{k0} = 0$) or odds ($\alpha_{k0} = 1$) assumption and (ii) H_0: $\beta_{kp} = 0$, $k = 1, \ldots, n_K$, $p = 1, \ldots, P$ (where $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kP})^T$), for testing the covariate effect on the cumulative incidence of type k events. To test (i), the statistic

$$Z_{\alpha_k} = \frac{\widehat{\alpha_k} - \alpha_{k0}}{\mathrm{SE}(\widehat{\alpha_k})}, \quad k = 1, \ldots, n_K, \tag{5.3}$$

may be used. To test the hypothesis (ii), we consider

$$Z_{\beta_{kp}} = \frac{\widehat{\beta_{kp}}}{\mathrm{SE}(\widehat{\beta_{kp}})}, \quad p = 1, \ldots, P, \;\; k = 1, \ldots, n_K. \tag{5.4}$$

From the theory of maximum likelihood estimation, the statistics Z_{α_k} and $Z_{\beta_{kp}}$ follow the standard normal distribution asymptotically when the corresponding null hypothesis holds. Likelihood ratio and score tests may also be utilized for testing.

6. Application to breast cancer data

The data comes from Protocol B-19, one of the earliest clinical trials on breast cancer treatment at NSABP. Two adjuvant chemotherapy regimens, methotrexate and 5-fluorouracil (MF) and cyclophosphamide plus MF (CMF), were compared among breast cancer patients with negative axillary lymph nodes and negative estrogen receptors. The patients have been followed 15+ years for cancer recurrence and mortality. Fisher *and others* (2004) reported an analysis of the 13-year update of the B-19 data. In this paper, we use a cohort of 1017 eligible patients with known pathological tumor sizes (510 in the MF arm; 507 in the CMF arm).

In the analysis, we define recurrence as any breast cancer recurrence in local, regional, or distant sites as first events. Other competing first events include second primary cancer other than breast and deaths without evidence of any disease. This definition implies disease-specific versusnon-disease-specific events in terms of breast cancer. In this cohort, the numbers of recurrences and other events are 211 and 161, respectively.

The focus of the analysis is the effects of treatment and other baseline prognostic factors on the cumulative incidence functions of recurrence and other events, which quantify long-term disease burden. Since all patients on NSABP B-19 are axillary lymph node negative and estrogen receptor negative, we only consider treatment group (trt), tumor size (tsize), and age at randomization (age) as potential covariates in our models. We analyze two models: a simple model with one covariate for the treatment effect and a full model based on the three covariates.

We begin by analyzing the model with a single treatment effect covariate (trt), coded 0 for the MF group and 1 for the CMF group. For comparison, the method of Fine and Gray (1999), denoted as F–G, was used to fit a semiparametric transformation model with $g_k(\upsilon_k) = \log\{-\log(1-\upsilon_k)\}$ in (3.3) and $u_k(\cdot)$ completely unspecified. Parametric generalized odds rate regression models with Gompertz base distribution were fit with $\alpha_k = 0$ (proportional hazards; PH$^{(G)}$), $\alpha_k = 1$ (proportional odds; PO$^{(G)}$), and with α_k estimated (GOR$^{(G)}$). Table 1 summarizes parametric and semiparametric estimates of the regression models for breast cancer recurrence and other competing events.

The negative sign of $\hat{\rho}$ for recurrence indicates that the estimated distribution is improper. Testing the null hypothesis H_0: $\rho \geqslant 0$ in this case gives a one-sided p-value less than 0.0001. Interestingly, for other events $\widehat{\rho} > 0$ and testing the null hypothesis H_0: $\rho \geqslant 0$ gives a one-sided p-value of 0.20, indicating the estimated cumulative incidence function is proper. These results make intuitive sense if one considers that the cumulative incidence functions are estimated using data from the observation period of Protocol B-19, which only spans 14 years. Few breast cancer recurrences occur after 10 years and the cumulative incidence function for recurrence plateaus between 10 and 14 years; see Figure 1. An improper distribution is clearly warranted. On the other hand, other events occur at a fairly steady rate over the entire time period, with the cumulative incidence increasing linearly up to 14 years; see Figure 1. While one would expect that this incidence curve would plateau at later times, over the first 14 years, the curve is better described by a proper distribution. Of course, it would be inappropriate to extrapolate the fitted Gompertz model to estimate long-term probabilities of other events.

The proportional hazards and proportional odds assumptions were tested using the statistic (5.3). The p-values from testing the proportional hazards (odds) assumptions are 0.41 (0.16) and 0.45 (0.46) for recurrence and other events, respectively. Neither of the models is rejected for either event type, reflecting the large variances of $\widehat{\alpha}$.

Table 1. *Parameter estimates (standard errors) from parametric and semiparametric regression models for cumulative incidence of recurrence (R) and other events (O) with trt as covariate in NSABP B-19 data; GOR$^{(G)}$ = parametric generalized odds rate model with Gompertz baseline, F–G = Fine and Gray (semiparametric) model, PO$^{(G)}$ = proportional odds model with Gompertz baseline, PH$^{(G)}$ = proportional hazards model with Gompertz baseline*

K	Model	$\widehat{\rho}$	$\widehat{\tau}$	$\widehat{\alpha}$	$\widehat{\beta}_{\text{trt}}$
R	GOR$^{(G)}$	−0.29 (0.05)	0.07 (0.01)	−1.40 (1.69)	−0.45 (0.17)
	F–G	—	—	—	−0.54 (0.14)
	PO$^{(G)}$	−0.23 (0.02)	0.08 (0.01)	1.00 (fixed)	−0.60 (0.16)
	PH$^{(G)}$	−0.25 (0.02)	0.08 (0.01)	0.00 (fixed)	−0.54 (0.14)
O	GOR$^{(G)}$	0.48 (0.58)	0.008 (0.004)	33.90 (44.80)	−0.32 (0.52)
	F–G	—	—	—	−0.10 (0.16)
	PO$^{(G)}$	0.05 (0.02)	0.010 (0.002)	1.00 (fixed)	−0.12 (0.17)
	PH$^{(G)}$	0.04 (0.02)	0.010 (0.002)	0.00 (fixed)	−0.10 (0.16)

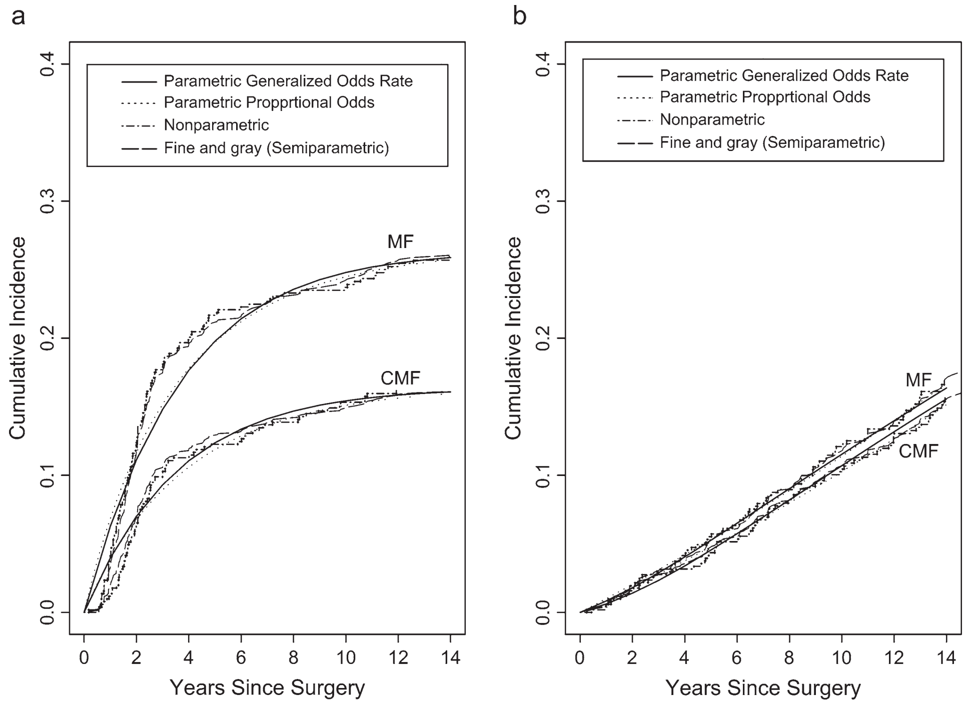

Fig. 1. Comparison of estimated cumulative incidence functions. (a) Recurrence, (b) other events.

Given the weak evidence against proportional hazards and odds models, we now consider analyses fixing $\alpha = 0$ or 1, which may have greater efficiency and greater interpretability than analyses in which α is estimated (see Table 1). Under the proportional hazards model, Wald tests based on the F–G semiparametric estimates of the treatment effects give p-values of 0.0001 for recurrences and 0.52 for the other events. After fixing $\alpha_k = 1$ under the proportional odds model with Gompertz baseline, Wald tests for the treatment effects result in p-values of 0.0001 and 0.50, for recurrence and other events. Under the parametric proportional hazards model with Gompertz baseline and $\alpha_k = 0$, the corresponding p-values are 0.0002 and 0.51. The estimated treatment effects under the semiparametric and parametric proportional hazards models are almost identical. Under the proportional odds model, the odds of recurrence on CMF is $\exp(-0.6) = 0.55$ that on MF, at all times t, based on the cumulative incidence functions. This interpretation may be more natural than that based on the proportional hazards model with $\alpha = 0$, where the regression parameters denote subdistribution hazard ratios.

Figure 1(a) shows a comparative plot of the nonparametric (dotted and dashed) estimates, semiparametric estimates from the proportional hazards model (dashed), and parametric estimates from the generalized odds rate model (solid) and the proportional odds model (dotted) of the cumulative incidence curves for recurrence up to year 14. Figure 1(b) shows a similar plot for the other events. To check the baseline Gompertz assumption, the parametric and semiparametric estimates of the baseline cumulative hazard functions were compared under the proportional hazards model (Figure 2). The parametric estimates of $u_0(t)$ are calculated as $u_0^{(\mathrm{PH}^{(G)})}(t; \widehat{\rho}, \widehat{\tau}) = \widehat{\tau}\{\exp(\widehat{\rho}t){-}1\}/\widehat{\rho}$, where $\widehat{\rho}$ and $\widehat{\tau}$ are the estimates from the parametric proportional hazards model with Gompertz baseline. The semiparametric estimates are calculated using $u_0^{(\mathrm{F-G})}(t) = -\log\{1 - \widehat{F}_{k0}^{(\mathrm{F-G})}(t)\}$, where $\widehat{F}_{k0}^{(\mathrm{F-G})}(t)$ is the semiparametric estimate of the baseline cumulative subdistribution from the F–G model. In both Figures 1 and 2, the parametric curves agree reasonably well with the nonparametric and semiparametric estimates, although there is

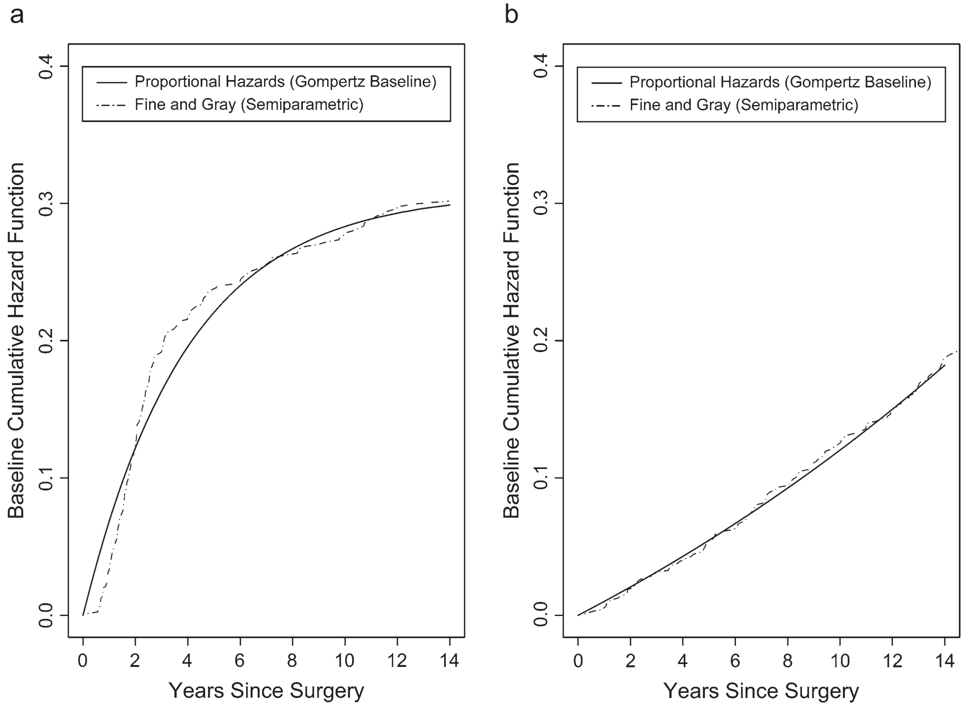

Fig. 2. Checking baseline distributional assumptions of Gompertz in NSABP B-19 data set. (a) MF group, (b) CMF group.

some evidence of lack of fit in the first few years of the follow-up period for breast cancer recurrence in Figure 1(a). There do not appear to be substantial differences between the different parametric fits. These findings suggest that the cumulative incidence functions may be well approximated by simple Gompertz models. Moreover, the estimated long-term failure probabilities are fairly insensitive to ρ in the odds rate model.

Next, we fit a regression model with age and tsize, as well as trt. The parametric and semiparametric estimates are summarized in Table 2. P-values for testing the proportional hazards (odds) assumptions are 0.73 (0.32) and 0.36 (0.27) for recurrence and other events, respectively. Again, there is large variability in estimation of α, which makes it difficult to differentiate between different transformations in (3.1).

For breast cancer recurrence, Wald tests for β_{trt}, β_{age}, and β_{tsize} give p-values of 0.0001, 0.321, and 0.010 from the parametric proportional odds model, 0.0001, 0.330, and 0.004 from the semiparametric proportional hazards model, and 0.0001, 0.335, and 0.012 from the parametric proportional hazards model with Gompertz baseline. For other non-breast-cancer-related events, p-values for the regression coefficients are 0.548, 0.003, and 0.839 from the parametric proportional odds model, 0.540, 0.008, and 0.930 from the F–G model, and 0.557, 0.003, and 0.808 from the parametric proportional hazards model with Gompertz baseline.

For both recurrence and other events, the results are rather consistent across models. There are significant treatment and tumor size effects on recurrence among node-negative and estrogen receptor-negative breast cancer patients in Protocol B-19. The observation of a positive correlation between tumor size and breast cancer recurrence rate is anticipated, but is still important information for patients and investigators. The decreased risk of recurrence with CMF persists after adjusting for initial disease severity, as measured by tumor size. For non-breast-cancer-related events, only the effect of age is significant.

Table 2. *Parameter estimates (standard errors) from parametric and semiparametric regression models for cumulative incidence of recurrence (R) and other event types (O), including trt, age, and tsize as covariates in NSABP B-19 data; GOR$^{(G)}$ = parametric generalized odds rate model with Gompertz baseline, F–G = Fine and Gray (semiparametric) model, PO$^{(G)}$ = proportional odds model with Gompertz baseline, PH$^{(G)}$ = proportional hazards model with Gompertz baseline*

K	Model	$\widehat{\rho}$	$\widehat{\tau}$	$\widehat{\alpha}$	$\widehat{\beta}_{\text{trt}}$	$\widehat{\beta}_{\text{age}}$	$\widehat{\beta}_{\text{tsize}}$
R	GOR$^{(G)}$	−0.27 (0.05)	0.08 (0.03)	−0.52 (1.54)	−0.51 (0.17)	−0.007 (0.007)	0.011 (0.006)
	F–G	—	—	—	−0.56 (0.14)	−0.007 (0.007)	0.013 (0.004)
	PO$^{(G)}$	−0.23 (0.02)	0.08 (0.03)	1.00 (fixed)	−0.61 (0.16)	−0.008 (0.008)	0.015 (0.006)
	PH$^{(G)}$	−0.25 (0.02)	0.08 (0.03)	0.00 (fixed)	−0.55 (0.14)	−0.007 (0.007)	0.013 (0.005)
O	GOR$^{(G)}$	−0.03 (0.08)	0.003 (0.004)	−4.78 (5.24)	−0.06 (0.10)	0.02 (0.01)	0.001 (0.003)
	F–G	—	—	—	−0.10 (0.16)	0.02 (0.01)	0.001 (0.007)
	PO$^{(G)}$	0.05 (0.02)	0.003 (0.001)	1.00 (fixed)	−0.10 (0.17)	0.03 (0.01)	0.001 (0.006)
	PH$^{(G)}$	0.04 (0.02)	0.003 (0.001)	0.00 (fixed)	−0.09 (0.16)	0.02 (0.01)	0.001 (0.006)

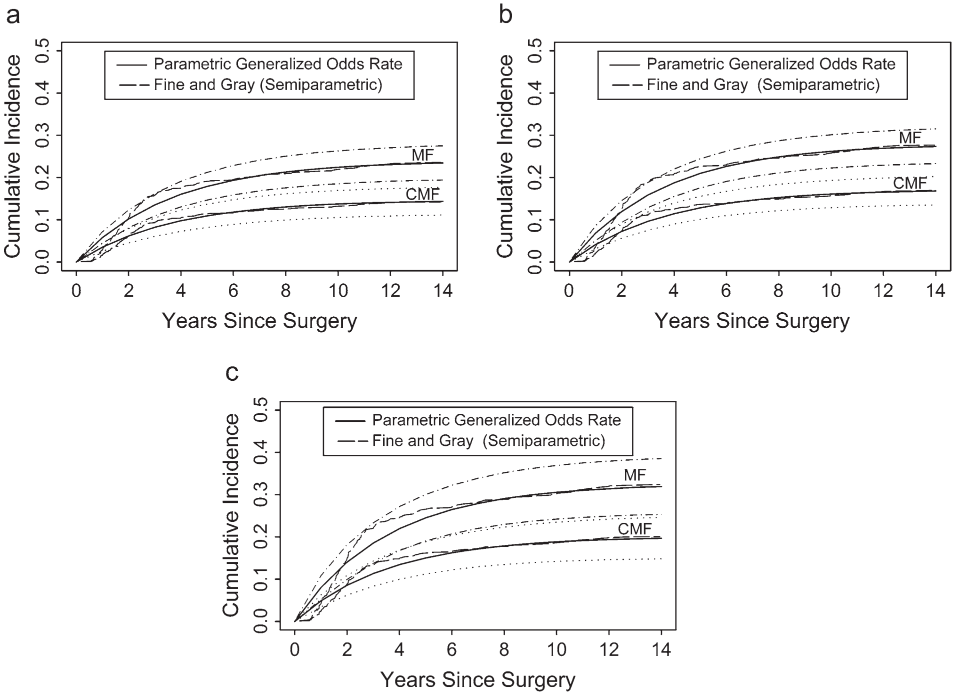

Fig. 3. Comparison of predicted cumulative incidence functions at different tumor sizes. (a) Tumor size = 15 mm, (b) tumor size = 30 mm, (c) tumor size = 45 mm.

This may be because the etiology of the other causes of failure is connected to the aging process, which is unaffected by either treatment or the size of the original breast tumor.

We consider estimation of the recurrence rates conditionally on treatment group and tumor size, whose effects are statistically significant in all models. The parametric generalized odds rate model with unknown α (smooth lines) and the semiparametric F–G model (dashed lines with jumps) are shown in Figure 3 for different values of tumor size in each treatment group. For the odds rate model, the point estimates are supplemented by 95% pointwise confidence intervals. The parametric curves are calculated with

$$\widehat{F}_1(t;\, \text{trt}, \text{tsize}) = 1 - [1 + \widehat{\alpha}\widehat{\tau}\exp(\widehat{\beta}_{\text{trt}} \times \text{trt} + \widehat{\beta}_{\text{tsize}} \times \text{tsize})\{\exp(\widehat{\rho}t) - 1\}/\widehat{\rho}]^{-1/\widehat{\alpha}},$$

where group = 0 for the MF arm or 1 for the CMF arm and tsize = 15, 30, or 45 mm. The two cumulative incidence estimates agree, on average, with the parametric model tending to underestimate at early time points. The semiparametric curves are within the pointwise confidence intervals from the parametric estimates, except for the first few years of the follow-up period. In each treatment group, the probability of recurrence increases as tumor size increases. Parametric estimates of 5- and 10-year recurrence rates in the MF arm at tumor sizes of 15, 30, and 45 mm are 17.9, 21.0, and 24.5% and 22.4, 26.2, and 30.6%, respectively.

The proportion of breast cancer patients never experiencing breast cancer recurrence, which can be interpreted as a cure fraction, can be estimated by replacing the parameters in (3.10) with their maximum likelihood estimates. Under the generalized odds rate model, this yields

$$1 - \widehat{F}_1(\infty;\, \text{trt}, \text{tsize}) = \{1 - \widehat{\alpha}\widehat{\tau}\exp(\widehat{\beta}_{\text{trt}} \times \text{trt} + \widehat{\beta}_{\text{tsize}} \times \text{tsize})/\widehat{\rho}\}^{-1/\widehat{\alpha}}.$$

For tumor size 20 mm (median in Protocol B-19), the estimated long-term cure rates are 74.7 and 84.4% for patients on MF and CMF, respectively. This represents a meaningful decrease in disease burden for breast cancer patients receiving CMF. It is important to recognize that such estimation is not possible under nonparametric and semiparametric models, where $u_k(t)$ cannot be estimated beyond the largest observed follow-up time.

7. DISCUSSION

In this paper, we proposed maximum likelihood inferences for a direct parametric regression modeling framework for the cumulative incidence function with competing risks data. A general parametric form of the transformation $g_k(v_k; \alpha_k)$ was considered, which includes proportional hazards and proportional odds models. The baseline distribution was modeled using a Gompertz specification, which accommodates improper distributions, like the cumulative incidence function. In theory, any parametric model could be used for $u_k(t)$, with the maximum likelihood estimators giving valid inferences under the usual regularity conditions.

The parametric analysis enables estimation of the long-term proportion of individuals experiencing a particular event type. This differs from standard nonparametric and semiparametric analyses, where $u_k(t)$ is completely unspecified. Of course, care should be exercised when interpreting parametric extrapolations, particularly when there is no evidence of a plateau in the tail of the estimated cumulative incidence function over the observed time period.

A two-parameter Gompertz distribution was used to parameterize the baseline distribution. It is worth noting that this model only allows monotone hazard shapes. Greater flexibility may be obtained with other parametric models which permit unimodal and bathtub shapes. The additional flexibility may be helpful in obtaining more accurate predictions of the cumulative incidence functions over the entire follow-up period, as in Protocol B-19, where there is some evidence of lack of fit at early time points. There may be other applications where the Gompertz model is seriously deficient and different specifications may be needed to obtain reliable results.

ACKNOWLEDGMENTS

This research was supported in part by the Department of Defense grant W81XWH-04-1-0605 and National Health Institute grants 5-U10-CA69974-09 and 5-U10-CA69651-11. *Conflict of Interest:* None declared.

References

AALEN, O. O. (1978). Nonparametric inference for a family of counting processes. *Annals of Statistics* **6**, 701–26.

BENICHOU, J. AND GAIL, M. H. (1990). Estimates of absolute cause-specific risk in cohort studies. *Biometrics* **46**, 813–26.

BOAG, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society, Series B, Methodological* **11**, 15–44.

BRYANT, J. AND DIGNAM, J. J. (2004). Semiparametric models for cumulative incidence functions. *Biometrics* **60**, 182–90.

COX, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological* **34**, 187–202.

COX, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269–76.

DABROWSKA, D. M. AND DOKSUM, K. A. (1998). Estimation and testing in a two-sample generalized odds-rate model. *Journal of the American Statistical Association* **83**, 744–9.

FINE, J. P. (2001). Regression modelling of competing crude failure probabilities. *Biostatistics* **2**, 85–97.

FINE, J. P. AND GRAY, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association* **94**, 496–509.

FISHER, B., JEONG, J., ANDERSON, S. AND WOLMARK, N. (2004). Treatment of lymph node-negative, estrogen receptor-negative breast cancer: updated findings from National Surgical Adjuvant Breast and Bowel Project clinical trials. *Journal of the National Cancer Institute* **96**, 1823–31.

FISHER, B., REDMOND, C., DIMITROV, N. V., BOWMAN, D., LEGAULT-POISSON, S., WICKERHAM, D. L., WOLMARK, N., FISHER, E. R., MARGOLESE, R., SUTHERLAND, C. *and others* (1989). A randomized clinical trial evaluating sequential methotrexate and fluorouracil in the treatment of patients with node-negative breast cancer who have estrogen-receptor-negative tumors. *The New England Journal of Medicine* **320**, 473–8.

GAYNOR, J. J., FEUER, E. J., TAN, C. C., WU, D. H., LITTLE, C. R., STRAUS, D. J., CLARKSON, B. D. AND BRENNAN, M. F. (1993). On the use of cause-specific failure and conditional failure probabilities: examples from clinical oncology data. *Journal of the American Statistical Association* **88**, 400–9.

GOMPERTZ, B. (1825) On the nature of the function expressive of the law of human mortality, and on the new mode of determining the value of life contingencies. *Philosophical Transactions of the Royal Society of London A* **115**, 513–80.

GRAY, R. J. (1988). A class of K-sample tests for comparing the cumulative incidence of a competing risk. *Annals of Statistics* **16**, 1141–54.

JEONG, J. AND FINE, J. P. (2006). Direct parametric inference for cumulative incidence function. *Journal of the Royal Statistical Society, Series C, Applied Statistics* **55**, 187–200.

JEONG, J. AND OAKES, D. (2003). On the asymptotic relative efficiency of estimates from Cox's model. *Sankhya* **65**, 411–21.

KALBFLEISCH, J. D. AND PRENTICE, R. L. (1980). *The Statical Analysis of Failure Time Data*. New York: John Wiley & Sons, Inc.

KAPLAN, E. L. AND MEIER, P. (1958). Nonparametric estimator from incomplete observations. *Journal of the American Statistical Association* **53**, 457–81.

KARRISON, T. G., FERGUSON, D. J. AND MEIER, P. (1999). Dormancy of mammary carcinoma after mastectomy. *Journal of the National Cancer Institute* **191**, 80–5.

KORN, E. L. AND DOREY, F. J. (1992). Applications of crude incidence curves. *Statistics in Medicine* **11**, 813–29.

LARSON, M. G. AND DINSE, G. E. (1985). A mixture model for the regression analysis of competing risks data. *Journal of the Royal Statistical Society, Series C, Applied Statistics* **34**, 201–11.

MEIER, P., KARRISON, T., CHAPPELL, R. AND XIE, H. (2004). The price of Kaplan–Meier. *Journal of the American Statistical Association* **99**, 890–6.

MILLER, JR, R. G. (1983). What price Kaplan–Meier? *Biometrics* **39**, 1077–81.

NELSON, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics* **14**, 945–65.

PEPE, M. S. (1991). Inference for events with dependent risks in multiple endpoint studies. *Journal of the American Statistical Association* **86**, 770–8.

PEPE, M. S. AND MORI, M. (1993). Kaplan–Meier, marginal or conditional probability curves in summarizing competing risks failure time data? *Statistics in Medicine* **12**, 737–51.

PRENTICE, R. L., KALBFLEISCH, J. D., PETERSON, JR, A. V., FLOURNOY, N., FAREWELL, V. T. AND BRESLOW, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics* **34**, 541–54.

TAGHIAN, A., JEONG, J., MAMOUNAS, E., ANDERSON, S., BRYANT, J., DEUTSCH, M. AND WOLMARK, N. (2004). Pattern of loco-regional failure in patients with breast cancer treated by mastectomy and chemotherapy (+/− tamoxifen) without radiation: results from five NSABP randomized trials. *Journal of Clinical Oncology* **22**, 4247–54.

[*Received January 12, 2006; revised April 19, 2006; accepted for publication April 21, 2006*]