

# Adjustment for Competing Risk in Kin-Cohort Estimation

Nilanjan Chatterjee,\* Patricia Hartge, and Sholom Wacholder

*Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland*

Kin-cohort design can be used to study the effect of a genetic mutation on the risk of multiple events, using the same study. In this design, the outcome data consist of the event history of the relatives of a sample of genotyped subjects. Existing methods for kin-cohort estimation allow estimation of the risk of one event at a time with the assumption that the censoring events are unrelated to the genetic mutation under study. These methods, however, may produce biased estimates of risk when multiple events are related to the genetic mutation, and follow-up of some of the events may be censored by the onset of other events. Using a competing risk framework to address this problem, we show that cause-specific hazard functions for carriers and noncarriers are identifiable from kin-cohort data. For estimation, we propose an extension of a composite-likelihood approach we described previously. We illustrate the use of the proposed method for estimation of the risk of ovarian cancer from BRCA1/2 mutations in the absence of breast cancer, based on data from the Washington Ashkenazi Kin-Cohort Study. We also evaluate the performance of the proposed estimation method, based on simulated data that were generated following the setup of the Washington Ashkenazi Study. *Genet Epidemiol* 25:303–313. Published 2003 Wiley-Liss, Inc.†

**Key words:** BRCA1/2 mutations; cause-specific hazard; cohort study; multiple outcomes; penetrance

\*Correspondence to: Nilanjan Chatterjee, Division of Cancer Epidemiology and Genetics, National Cancer Institute, 6120 Executive Blvd, EPS Room 8038, Rockville, MD 20852. E-mail: chattern@mail.nih.gov

Received 18 March 2003; Accepted 16 June 2003

Published online in Wiley InterScience (www.interscience.wiley.com)

DOI:10.1002/gepi.10269

## INTRODUCTION

Struewing et al. [1997] first introduced the kin-cohort design in the context of the Washington Ashkenazi Study (WAS). In this study, blood sample and questionnaire data were collected from 5,318 Ashkenazi Jewish men and women volunteers (probands) living in the Washington, DC area. Based on blood samples, volunteers were tested for three specific founder mutations for this population in BRCA1/2 genes. Struewing et al. (1997) estimated the absolute risk of breast cancer (penetrance) among BRCA1/2 mutation carriers, using family history data of the genotyped volunteers. Wacholder et al. [1998], who formally proposed this novel approach as a kin-cohort design, established the main analytic principle behind this method, showing how penetrance can be estimated by relating the disease history of the relatives to the genotypes of the probands. Wacholder et al. [1998] and Gail et al. [1999b] discussed various practical advantages of this approach, as well as some of its limitations. Gail et al. [1999a], Moore et al. [2001], and Chatterjee and Wacholder [2001] described various extensions of

the original analytic approach of Wacholder et al. [1998] for estimating age-specific penetrance (cumulative risk) of a disease from kin-cohort data.

A known advantage of the kin-cohort design is its ability to study multiple outcomes, using retrospective cohort data from the relatives. Although prospective cohort studies can be used to collect similar data in principle, the logistics and time needed for implementing such a design can be quite daunting relative to the kin-cohort design. The data from the Washington Ashkenazi Kin-Cohort Study, for example, were already utilized to examine the effect of BRCA1/2 mutations on the risks of several different cancers [Struewing et al., 1997] and on survival after the onset of breast cancer [Lee et al., 1999].

Currently, we are analyzing the mortality history data of WAS relatives to explore the association between BRCA1/2 mutations and the risk of overall mortality. Since BRCA1/2 mutations are well-known to be related to a major increase in the risk of breast and ovarian cancer in women, and possibly prostate cancer in men, we are mainly interested in examining possible associations between the

mutations and the risk of mortality that cannot be explained by deaths from these known BRCA1/2-related cancers. Thus, we consider mortality in the absence of these cancers as the primary endpoint of interest. Relatives who were diagnosed with these cancers are considered to be censored at the onset of their cancers. The main scientific results from this study will be presented elsewhere.

A common assumption for all of the existing methods for kin-cohort estimation [Wacholder et al., 1998; Gail et al., 1999a; Moore et al., 2001; Chatterjee and Wacholder, 2001] is that the censoring mechanism does not depend on the mutation under study. This assumption is clearly violated in our mortality study, because the onset of any of the known BRCA1/2-related cancers is treated as a censoring event. The same problem may arise more generally. In an evaluation of the risk of ovarian cancer, for example, subjects may be censored because death from breast cancer could occur before ovarian cancer that would otherwise be diagnosed. In this article, we study the methods for quantifying and estimating disease risk from kin-cohort data when the risk of censoring events may be related to the mutations. First, we review the original analytic approach of Wacholder et al. [1998] and examine why the independent censoring assumption was needed in that approach. Then we introduce a competing risk framework to address this problem. We show that cause-specific hazard functions, a concept widely used in the analysis of standard cohort data, can be identified in a kin-cohort setting. We discuss the interpretation of these functions and the conditions under which they can be translated to estimates of penetrance (cumulative risk) function. Next we propose the use of a composite-likelihood approach [Chatterjee and Wacholder, 2001] and a related Expectation-Maximization (EM) algorithm for the estimation of cause-specific hazard functions in carriers and noncarriers. We apply the proposed method to data from the WAS study, to estimate the risk of ovarian cancer from BRCA1/2 mutations in the absence of breast cancer. Finally, we evaluate the performance of the proposed method, based on simulated data.

## METHODS

### ORIGINAL ANALYTIC APPROACH FOR KIN-COHORT ESTIMATION

Let us assume we are studying a genetic mutation with a dominant mode of inheritance,

so that  $g = 0$  corresponds to a subject who does not carry the mutation, and  $g = 1$  corresponds to subjects who carry one or two copies of the mutation. Let  $F_g(t)$  denote the cumulative risk (penetrance) of a disease up to age  $t$  associated with genotype  $G = g$ . The penetrance functions  $F_0(t)$  and  $F_1(t)$  cannot be directly estimated from kin-cohort data, because the mutation status of the relatives is unknown. Wacholder et al. (1998) observed that, for a rare mutation with allele frequency  $f$ , the odds of carrying the mutation among first-degree relatives of noncarriers and carriers are given by  $2f : (1 - 2f)$  and  $(0.5 + f) : (0.5 - f)$ , respectively. Consequently, the cumulative risks of disease among first-degree relatives of noncarriers and carriers are given by the equations

$$\begin{aligned} R_0(t) &= (1 - 2f)F_0(t) + 2fF_1(t) \\ R_1(t) &= (1/2 - f)F_0(t) + (1/2 + f)F_1(t) \end{aligned} \quad (1)$$

respectively. Wacholder et al. [1998] observed that  $R_0(t)$  and  $R_1(t)$  in Equation (1) can be directly estimated using the Kaplan-Meier disease incidence curves for the relatives of noncarriers and the relatives of carriers, respectively. Thus if allele frequency  $f$  is known or can be externally estimated, then the two equations in (1) can be solved for each  $t$  to obtain an estimate of  $F_0(t)$  and  $F_1(t)$ .

In Equation (1), the cumulative risk of the disease for a person up to a given age  $t$  is defined, assuming the person is not censored from other causes before age  $t$ . In the presence of censoring, it is well-known that the Kaplan-Meier incidence curve gives an unbiased estimate for the cumulative risk function only if the risk of the disease and the risk of censoring events are independent. Specifically, Kaplan-Meier estimation of  $R_0(t)$  and  $R_1(t)$  will be valid only if the independent censoring assumption holds separately for the relatives of carriers and the relatives of noncarriers. Clearly, the risks of the disease and the censoring are correlated if both of the events are affected by the mutation under study. For the relatives of noncarriers, very little correlation will be induced by a rare mutation, since only a small fraction of these relatives will be carriers. By contrast, the mutation will be relatively common (about 50%) in the relatives of carriers, and thus the correlation induced will be important to consider. When the correlation is strongly positive, the Kaplan-Meier incidence curve for the relatives of carriers could seriously underestimate

$R_1(t)$ . Since the solution of  $F_1(t)$  from Equation (1) is given by  $2R_1(t) - R_0(t)$ , the resulting estimate of  $F_1(t)$  also will be an underestimate of the true cumulative risk function (see Fig. 1).

In an ordinary cohort study, where the mutation status of cohort members could be directly observed, the above issue does not arise. In this case, the cumulative risk of the disease for noncarriers and carriers can be directly estimated by the corresponding Kaplan-Meier disease incidence curves, using the weaker assumption that the risks of the disease and the censoring are independent, conditional on the mutation status of a subject.

**CAUSE-SPECIFIC HAZARD: DEFINITION, INTERPRETATION AND IDENTIFIABILITY**

For dealing with multiple events, where some of the events may censor the follow-up time for other events, a “competing risk” model provides the general framework. Let  $T_1$  and  $T_2$  denote time to two competing events  $E_1$  and  $E_2$ . We will assume the standard definition of “competing risks,” i.e., for any given subject, only the first of the two events is observable, and the follow-up for the

second event ends at the onset of the first event. In the presence of competing risks, it is useful to think of estimation in terms of “cause-specific hazard” functions. The cause-specific hazard function for the  $i$ th event at time  $t$  for an individual with genotype  $G = g$  can be defined as

$$\lambda_{ig}(t) = \lim_{\delta \downarrow 0} \frac{1}{\delta t} \Pr\{T_i \in [t, t + \delta t) | T_1 \geq t, T_2 \geq t, G = g\}. \tag{2}$$

That is,  $\lambda_{ig}(t)$  is the instantaneous probability that an individual with genotype  $G = g$  will experience the event  $E_i$  at time  $t$ , given that s/he has been “at risk,” i.e., has been free of both events until time  $t$ .

Before describing estimation, it is instructive to understand the identifiability of these cause-specific hazard functions. When dealing with a single event, Wacholder et al. [1998] established the identifiability of the genotype-specific cumulative risk functions by establishing their relationship with the cumulative disease incidence functions for the relatives of carriers and the relatives of noncarriers (see Equation 1). Next, we show that while dealing with multiple competing

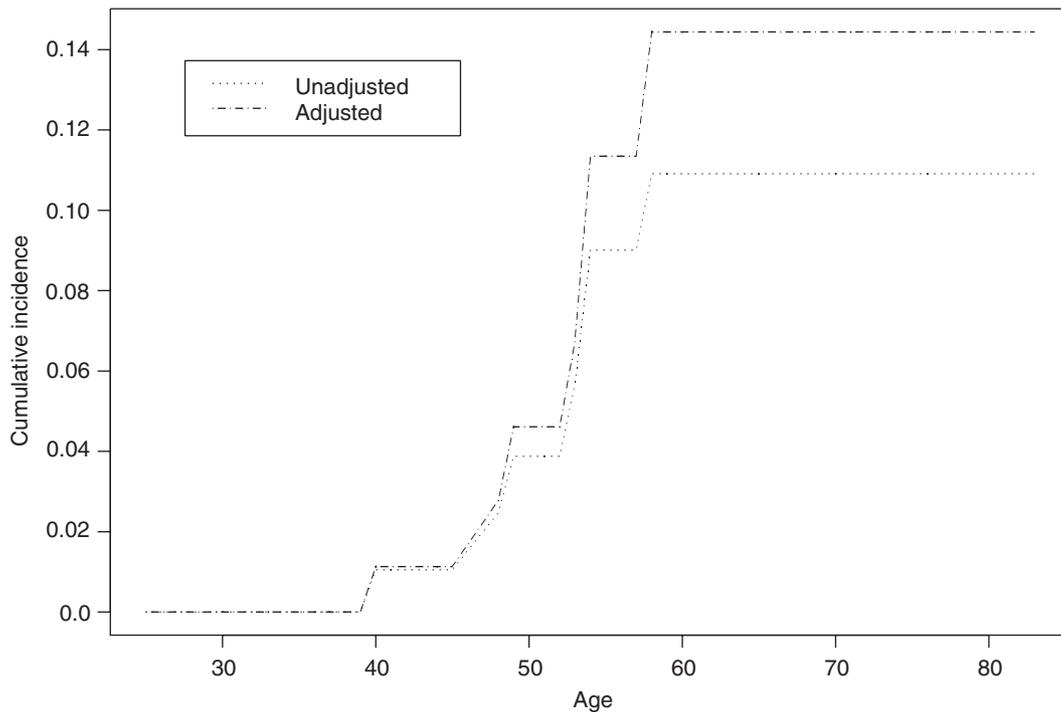


Fig. 1. Estimated age-specific cumulative incidence of first ovarian cancer (ovarian cancer in absence of breast cancer), based on Washington Ashkenazi Study. Dotted line shows estimate that ignores effect of BRCA1/2 mutations on breast cancer. Dashed line shows estimate that adjusts for effect of mutations on risk of breast cancer.

events, similar relationship can be established through cause-specific hazard functions.

Let us define

$$r_{ig}(t) = \lim_{\delta t \downarrow 0} \frac{1}{\delta t} \Pr\{T_i \in [t, t + \delta t) | T_1 \geq t, T_2 \geq t, G_0 = g\} \quad (3)$$

to be the cause-specific hazard function for the  $i$ th event among relatives of probands with genotype  $G_0 = g$ . Each of the four cause-specific hazard functions  $r_{ig}(t)$ ,  $i = 1, 2$ ;  $g = 0, 1$  can be empirically estimated by the corresponding proportion of "at-risk" relatives who experience the given event at time  $t$ . As derived in the Appendix,  $r_{ig}(t)$ ,  $i = 1, 2$ ;  $g = 0, 1$  now can be related to the cause-specific hazard functions of interest  $\lambda_{ig}(t)$ ,  $i = 1, 2$ ;  $g = 0, 1$  by the equations

$$r_{ig}(t) = \sum_{g'=0,1} \lambda_{ig'}(t) \Pr(G = g' | G_0 = g, T_1 \geq t, T_2 \geq t), \quad (4)$$

where

$$\Pr(G = g' | G_0 = g, T_1 \geq t, T_2 \geq t) = \frac{\Pr(T_1 \geq t, T_2 \geq t | G = g') \Pr(G = g' | G_0 = g)}{\sum_{g'=0,1} \Pr(T_1 \geq t, T_2 \geq t | G = g') \Pr(G = g' | G_0 = g)}.$$

Above, the joint probability  $\Pr(T_1 \geq t, T_2 \geq t | G = g')$  can be characterized by the cause-specific hazard functions of individual events, using the standard formula [e.g., Chapter 7 in Kalbfleisch and Prentice, 1978]:

$$\Pr(T_1 \geq t, T_2 \geq t | G = g') = \exp\left\{-\int_0^t \lambda_{1g'}(s) ds\right\} \exp\left\{-\int_0^t \lambda_{2g'}(s) ds\right\}.$$

Further,  $\Pr(G = g' | G_0 = g)$ , the conditional probability of a relative's genotype given that of the proband, can be computed as a function of the allele frequency, assuming a Mendelian mode of inheritance. Thus, if the allele frequency is known or can be externally estimated, the equations defined by (4) for  $i = 1, 2$  and  $g = 0, 1$  yield four equations in the four cause-specific hazard functions:  $\lambda_{ig}(t)$ ,  $i = 1, 2$ ;  $g = 0, 1$ . Since  $r_{ig}(t)$  in Equation (4) can be directly estimated from kin-cohort data, these are four equations in four unknowns, and hence have enough information to identify the unknown functions  $\lambda_{ig}(t)$ ,  $i = 1, 2$ ;  $g = 0, 1$ . However, the above arguments do not constitute a theoretical proof; they do provide the basic

intuition behind the identifiability of the cause-specific hazard functions.

Various transformations of cause-specific hazard functions are often of interest for the presentation of data. One such transformation is the "cumulative incidence function," which can be defined as

$$F_{ig}(t) = 1 - S_{ig}(t) = 1 - \exp\left\{-\int_0^t \lambda_{ig}(s) ds\right\}. \quad (5)$$

for event type  $i$  and genotype  $g$ . The cumulative incidence function corresponds to the Kaplan-Meier incidence curve in a standard survival analysis setting. It is often described as the age-specific penetrance function in the literature of genetic epidemiology. The interpretation of the cumulative incidence function as cumulative risk, however, requires some caution. First, it requires the assumption that the risks of the outcome of interest and the competing/censoring event are independent, conditional on the genetic mutation. With this assumption, the cumulative incidence function of an event can be interpreted as the cumulative risk for that event in the hypothetical population obtained by removing all competing events, under the assumption that the removal of competing events does not change the risk of the event of interest. Because the cumulative risk interpretation refers to a hypothetical state, typically it may be best to view the cumulative incidence function, not as a cumulative risk, but as a simple way of summarizing and standardizing cause-specific hazard rates. Such a summarization can be useful for the comparison of age-specific rates for different populations with different age-distributions or/and different rates of the competing events.

## ESTIMATION

In principle, the equations given in (4) can be iteratively solved to estimate the cause-specific hazard functions of interest. One oddity is that hazard estimates from these equations cannot be guaranteed to be non-negative. We propose a likelihood based estimation constrained to avoid such anomalies.

We employ piecewise constant modelling of the cause-specific hazard functions  $\lambda_{ig}(t)$ ,  $i = 1, 2$ ;  $g = 0, 1$ . For the  $i$ th event, let  $\{t_l^{(i)}\}_{l=0}^{k_i-1}$  denote a set of knots appropriately chosen in the range of  $T_i$  so that  $0 = t_0^{(i)} < t_1^{(i)} < t_2^{(i)} < \dots < t_{k_i}^{(i)} < t_{k_i+1}^{(i)} = \infty$ . We will assume that both  $\lambda_{i0}(t)$  (hazard for noncarriers) and  $\lambda_{i1}(t)$  (hazard for carriers) are piecewise,

constant within intervals defined by  $\{t_l^{(i)}\}_{l=0}^{k_i+1}$ . Let the hazard parameters for carriers and noncarriers in the corresponding  $k_i + 1$  intervals be denoted by  $\{\lambda_{gl}^{(i)}\}_{l=0}^{k_i+1}$ ,  $g = 0, 1$ , respectively. This flexible approach of piecewise constant modeling of hazard functions can be nonparametric, in essence, if we define the set of knot points for each event to be the set of time points where at least one event of the respective type has been observed.

For estimation of hazard parameters in the above model, we propose the use of a “composite likelihood” approach that we defined earlier [Chatterjee and Wacholder, 2001] for general kin-cohort estimation. Earlier we had used the term “marginal likelihood” instead of “composite likelihood,” but since then, several readers noted that the latter term was more accurate for our methodology. In this approach, the likelihood contribution of family history data of the relatives of a volunteer is computed as the product of the probabilities of the phenotype history of the individual relatives, given the genotype of the volunteers. The advantages of this approach compared to the true likelihood of the data [Gail et al., 1999a] that is based on the full joint-probability distribution of the event histories of the relatives in a family were described in Chatterjee and Wacholder [2001].

Some further notation is needed to define the composite-likelihood more formally. Let  $m$  denote the number of probands, and  $G_{0i}$  the genotype of the  $i$ th proband. Suppose the  $i$ th volunteer reports the family history of a phenotype  $Y$  for  $n_i$  relatives. Let  $Y_{ij}$  denote the value of  $Y$  for the  $j$ th relative of the  $i$ th proband. With these notations in mind, the composite-likelihood of the family history data of the relatives can be defined as

$$\prod_{i=1}^m \prod_{j=1}^{n_i} p(Y_{ij}|G_{0i}) = \prod_{i=1}^m \prod_{j=1}^{n_i} \sum_{g=0}^1 p(Y_{ij}|G_{ij} = g) \times Pr(G_{ij} = g|G_{0i}). \tag{6}$$

On the left-hand side of Equation (6),  $p(Y_{ij}|G_{0i})$  denotes the marginal probability density (or mass function) of the phenotype history of the  $j$ th relative of the  $i$ th proband, given the genotype of the  $i$ th proband. On the right-hand side, this probability is computed as the weighted sum of

the probability density of the phenotype history of the relative if the relative was a noncarrier ( $G_{ij} = 0$ ) and if the relative was a carrier ( $G_{ij} = 1$ ), with weights defined by the corresponding probabilities of the relative being a noncarrier and a carrier given the genotype of the volunteer. Although Equation (6) does not define the true likelihood of the data, the theoretical basis for its validity follows from the fact that the score equations (derivative of the log-likelihood) corresponding to Equation (6) give an unbiased estimating equation [Godambe, 1991].

With two competing events, the phenotype history  $Y$  of a relative can be represented by a triplet of observations  $Y = (T, \Delta_1, \Delta_2)$ . Here,  $T$  denotes the time to the first of the two events, or censoring if neither of the events occurred during follow-up, and  $\Delta_k, k = 1, 2$  denote the indicator of whether the events  $E_1$  and  $E_2$  occurred or not, respectively. Here  $\Delta_1 = 1$  and  $\Delta_2 = 1$  cannot occur simultaneously because of the competing risk framework. The composite-likelihood of the event history data of relatives can be computed by replacing  $p(Y_{ij}|G_{ij} = g)$  in Equation (6) with the corresponding likelihood for competing risk data (e.g., Chapter 7 in Kalbfleisch and Prentice, 1978), given by

$$\lambda_{1g}(T_{ij})^{\Delta_{1ij}} \lambda_{2g}(T_{ij})^{\Delta_{2ij}} S_{1g}(T_{ij}) S_{2g}(T_{ij})$$

where  $S_{ig}(t), i = 1, 2; g = 0, 1$  are defined in Equation (5).

We propose use of an EM algorithm for maximization of the composite-likelihood with respect to the hazard parameters. Although the EM algorithm is traditionally used for maximum-likelihood estimation in missing data problems, a similar algorithm can be more generally applied to any method that is based on unbiased estimating equations [e.g., Rosen et al., 2000]. In our application, the E-step of the algorithm involves computing the conditional probability of each relative being a carrier and a noncarrier, given their individual event history and the genotype of the index proband. Let  $w_{0ij}$  and  $w_{1ij}$  denote the corresponding probabilities of being a noncarrier and a carrier, respectively, for the  $j$ th relative of the  $i$ th proband. In each iteration of the EM algorithm, these probabilities can be estimated from the current estimate of the hazard functions, using the formula

$$w_{gij} = \frac{\lambda_{1g}(T_{ij})^{\Delta_{1ij}} \lambda_{2g}(T_{ij})^{\Delta_{2ij}} S_{1g}(T_{ij}) S_{2g}(T_{ij}) Pr(G_{ij} = g|G_{0i})}{\sum_{g'=0}^1 \lambda_{1g'}(T_{ij})^{\Delta_{1ij}} \lambda_{2g'}(T_{ij})^{\Delta_{2ij}} S_{1g'}(T_{ij}) S_{2g'}(T_{ij}) Pr(G_{ij} = g'|G_{0i})}. \tag{7}$$

The M-step of the algorithm obtains an estimate of the hazard parameters given by the closed-form formula

$$\hat{\lambda}_{gk}^{(l)} = \frac{\sum_{i,j} N_{ijk}^{(l)} w_{gij}}{\sum_{i,j} PY_{ijk}^{(l)} w_{gij}}; \quad g = 0, 1; \quad (8)$$

$$k = 1, \dots, K_l; \quad l = 1, 2$$

where  $PY_{ijk}^{(l)}$  denotes the number of person years the  $i$ th relative of the  $j$ th proband contributes to the age interval  $[t_k^{(l)}, t_{k+1}^{(l)})$ , and  $N_{ijk}^{(l)}$  denotes the indicator of whether or not the relative has an event of type  $l$  in that interval. Iterating between the E-step and the M-step of the algorithm, until convergence yields the final estimates of hazard parameters.

In other words, each step of the EM algorithm involves estimating the cause-specific hazard rates for different age intervals in carriers and noncarriers by using a "number of events per person-year" formula (Equation 8) that is commonly used for standard cohort analysis. The genotype of relatives being unknown, the events and person years corresponding to a relative cannot be assigned as a whole to either carrier or noncarrier. Instead, the events and person years for each relative are divided between carrier and noncarrier according to the conditional probability of the relative being a carrier and a noncarrier, given the event history of the relative and the genotype of the index proband. Computation of these conditional probabilities ( $w_{gij}$ ), as shown in Equation (7), involves the hazard parameters themselves. Thus, the final estimates of hazard parameters are obtained by iterative use of Equation (8), where at each iteration, the conditional probabilities ( $w_{gij}$ ) are updated using Equation (7), based on the estimates of hazard parameters from the previous iteration.

#### ASYMPTOTIC THEORY AND VARIANCE ESTIMATION

For parametric/fixed-knot piecewise exponential models, the consistency (asymptotic unbiasedness) of the composite-likelihood estimation method follows from standard estimating equation theory (Godambe, 1991). For nonparametric models that allow a knot at each observed event time, although similarly consistent results can be expected to hold, a rigorous proof is not yet available. In simulation experiments (below), when we evaluate the performance of the non-

parametric estimation method on simulated data, the method is found to perform very well, even for quite a small sample size (see Fig. 2). In practice, one can use the nonparametric method for exploratory purposes, and then select a suitable parametric model such as a piecewise exponential model with fixed knots for formal inference. In this approach, the adequacy of a parametric model can be tested by comparing estimates of the quantities of scientific interest from the parametric model against those from the nonparametric approach.

To assess uncertainty in the estimates, in principle, one can use estimating-equation-based variance estimators such as the so called robust-sandwich method that is widely used in the generalized estimating equation (GEE) literature (Liang et al., 1992). Although these methods known to work well for parametric models, their performance in a nonparametric setting has not been well-studied. In our data application, we used a bootstrap-based resampling method [Efron and Tibshirani, 1998] that is known to perform well for both parametric and nonparametric models. To account for possible familial correlation between the relatives of the same proband, we use families as bootstrap sampling units. If there are  $M$  unique families corresponding to  $M$  probands in the study, in each bootstrap sample, we draw  $M$  families with replacement from the total set of  $M$  families. Once a bootstrap sample of families is chosen, the proposed method is used to obtain bootstrap estimates of the parameters. The empirical percentiles for bootstrap estimates over different bootstrap samples are used to define the confidence intervals for the parameter estimates.

## DATA EXAMPLE

We consider an application of the proposed method, using data from the Washington Ashkenazi Study (WAS). Based on these data, we previously reported estimates of age-specific cumulative risk of ovarian cancer among BRCA1/2 mutation carriers and noncarriers [Struewing et al., 1997; Chatterjee and Wacholder, 2001]. In these analyses, it was assumed that the goal was to estimate the risk of any ovarian cancer, irrespective of whether the cancers followed a previous breast cancer or not. Thus, history of breast cancer among relatives was ignored. An alternative strategy for analysis of these data could be to estimate the risk of first ovarian

cancer, i.e., the risk of ovarian cancer in the absence of previous breast cancer. If the etiology of ovarian cancer changes after the onset of breast cancer, e.g., due to treatment or/and change in lifestyle, the risk of first ovarian cancer may be different from that of any ovarian cancer. Such a distinction in the interpretation of risk can be particularly important for BRCA1/2 mutation carriers, since a large fraction of them are expected to develop breast cancer in their lifetime.

Table I shows the number of breast and ovarian cancer cases among female first-degree relatives of relatives of the WAS participants. We estimated the risk of first ovarian and first breast cancer by considering breast and ovarian cancers to be censoring/competing events for each other. For both events, we used nonparametric piecewise exponential models for the hazard functions that allow one knot at each age value where at least one event of the respective type was observed. We estimated BRCA1/2 allele frequency as 0.0112, based on the genotype data of the study participants. We obtained confidence intervals (CI) for parameter estimates, based on 500 bootstrap samples.

Table II shows the cause-specific hazard estimates for ovarian cancer in BRCA1/2 mutation carriers in the absence of breast cancer. In Table II, the adjusted estimates are obtained by the proposed methodology that accounts for the fact that the competing risk of breast cancer is related to BRCA1/2 mutations. The unadjusted estimates, on the other hand, are obtained by ignoring the relationship between breast cancer and BRCA1/2 mutations. We observe that adjustment due to the

competing risk of breast cancer was important for interval risk estimation for both of the age categories 40–50 and 50–60. Figure 2 shows the corresponding estimated age-specific cumulative incidence function (see Equation 5 for definition). Figure 2 shows a substantial difference in the adjusted and unadjusted estimates of lifetime cumulative incidence. The adjusted estimate of the cumulative incidence function up to age 70 was 0.144 (95% CI, 0.041–0.226), about 33% larger than the corresponding unadjusted estimate of 0.109 (95% CI, 0.032–0.159). We note that the adjusted estimate of lifetime cumulative incidence of first ovarian cancer was very close to the corresponding cumulative incidence estimate of any ovarian cancer that was previously reported (Struewing et al., 1997), based on the same data. This suggests that in our data, the risk of ovarian cancer due to BRCA1/2 mutations was similar before and after the onset of breast cancer.

In the context of this example, insight may be obtained by consideration of the proper interpretation of cumulative incidence. We estimated the lifetime (up to age 70) cumulative incidence of ovarian cancer among BRCA1/2 carriers to be about 14.4%. This estimate summarizes the age-specific rate of ovarian cancer in the absence of prior breast cancer among BRCA1/2 mutation carriers, and is useful for comparing our results with other studies. Interpretation of the estimate of the cumulative incidence as a cumulative risk, however, requires the additional assumption that the risks of breast and ovarian cancer are independent of each other among carriers of BRCA1/2 mutations. Even when the assumption holds, the estimate of 14.4% would correspond to the cumulative risk of ovarian cancer for carriers only in the hypothetical state in which carriers have no risk of breast cancer, either from BRCA1/2 mutations or from other causes. Given that the risk of breast cancer is very high for BRCA1/2 mutation carriers, the cumulative risk estimate could be directly applicable only for women who have some kind of intervention (e.g., bilateral mastectomy) that removes most or all of the risk of breast cancer but does not change the risk of ovarian cancer. Due to this restricted interpretation, cumulative incidence function should not typically be used for clinical risk prediction, for which other measures such as cumulative incidence in the presence of competing causes [Prentice et al., 1978; Gail et al., 1989] can be more useful.

**TABLE I. Distribution of breast cancer and ovarian cancer cases in relatives of WAS participants<sup>a</sup>**

	Any BC	First BC	Any OC	First OC
Relative of noncarriers (N=12,980)	982	976	119	111
Relatives of carriers (N=305)	58	58	11	7

<sup>a</sup>BC, breast cancer; OC, ovarian cancer.

**TABLE II. Cause-specific hazard of ovarian cancer in carriers of BRCA1/2 mutations integrated in three age intervals**

	Unadjusted estimate (95% CI)	Adjusted estimate (95% CI)
≤40	0.011 (0.000, 0.028)	0.011 (0.000, 0.026)
40–50	0.029 (0.000, 0.061)	0.036 (0.000, 0.080)
50–60	0.075 (0.000, 0.127)	0.107 (0.000, 0.190)

## SIMULATION EXPERIMENTS

### ESTIMATING RISK OF OVARIAN CANCER IN ABSENCE OF BREAST CANCER

We use simulation experiments to evaluate the performance of the proposed composite-likelihood method for estimation of the true (known) hazard and the cumulative incidence functions. In the first simulation, we generated data in a setting similar to the data application we described in the data example (above). We use an allele frequency of 0.0112 to generate the mutation status for 5,000 probands. Given the mutation status of the probands, we generate the mutation status for two first-degree relatives, say a mother and a sister, based on a Mendelian mode of inheritance. We assume that the mutation status of the relatives is unknown during analysis of the data. Times to onset of breast and ovarian cancers for the relatives are generated from Weibull distributions, using parameter values such that the cumulative risks of these diseases until ages 50 and 70 correspond to those we reported previously [Chatterjee and Wacholder, 2001]. Specifically, for describing breast cancer risk, we choose the shape and the scale parameter of the Weibull distribution to be 0.0078 and 3.2893, respectively, for noncarriers, and 0.0130 and 2.1334, respectively, for carriers. For describing ovarian cancer risk, we chose the corresponding shape and the scale parameters to be 0.0051 and 4.0051, respectively, for noncarriers, and 0.0081 and 2.9837, respectively, for carriers. Following the mechanism of censoring in the Washington Ashkenazi Study, we assume that relatives can be censored either at their death from other causes, or at the time of the interview of the proband. For both carrier and noncarrier relatives, we generate age at mortality from a normal distribution that has a mean age of 81 and standard deviation of 10. We generate current age (age at the time of the interview of the proband) for relatives using normal distribution, with mean age 70 for mothers and 50 for sisters, and a common standard deviation of 10.

We simulate 100 data sets in the above setting. Similar to the data example (above), we assume that our goal is to estimate the risk of ovarian cancer in the absence of breast cancer. Thus, we treat onset of breast and ovarian cancer as censoring/competing events for each other. We analyzed each data set, using the nonparametric version of the piecewise exponential hazard

model that allows one knot at each age value where at least one event of the respective type is observed in the data. We assume that the allele frequency is known.

Figure 2 shows estimate and true values for the cause-specific hazard and cumulative incidence functions for ovarian cancer in the carriers of the mutation. The solid line in Figure 2 shows the true (known) hazard/cumulative-incidence function corresponding to the underlying Weibull distribution for the carriers. The dashed line in Figure 2 shows the corresponding mean of the nonparametric estimates of the hazard/cumulative incidence function, adjusted for competing risk using the methods developed in this paper. The dotted line in Figure 2 shows the estimate of the corresponding functions, ignoring the effect of the mutation on the competing event of breast cancer. As the nonparametric estimates of hazard functions tend to be very irregular (discontinuous), we plot the hazard estimates after smoothing them, using a moving average method that allows for a 10-year window.

From Figure 2, we observe that ignoring the effect of the mutation on competing risk caused a very significant bias for both the hazard and the cumulative incidence estimation. For both of these functions, the bias seems to be more important for older ages than for younger ages. The proposed method, which adjusts for competing risk, on average estimates the true cumulative risk and hazard functions with very minimal bias. Given the fact that in our setting the expected number of ovarian cancer cases in the relatives of carriers is quite small (typically less than 10), our nonparametric estimation method seems to perform quite well, even with a small sample size.

### ESTIMATING RISK OF MORTALITY IN ABSENCE OF BREAST CANCER

Since the research is originally motivated by the goal of estimating the risk of mortality from BRCA1/2 mutations in the absence of known BRCA1/2-related cancers, we considered a second simulation study where we allowed the risk of mortality from other causes to be associated with the mutation. We generate data using the exact same setup as above, except that now we assume the mean age at mortality from other causes to be smaller for carriers than for noncarriers (73 vs. 81). Since estimating risk of mortality in the absence of breast cancer is of interest, we treat breast cancer

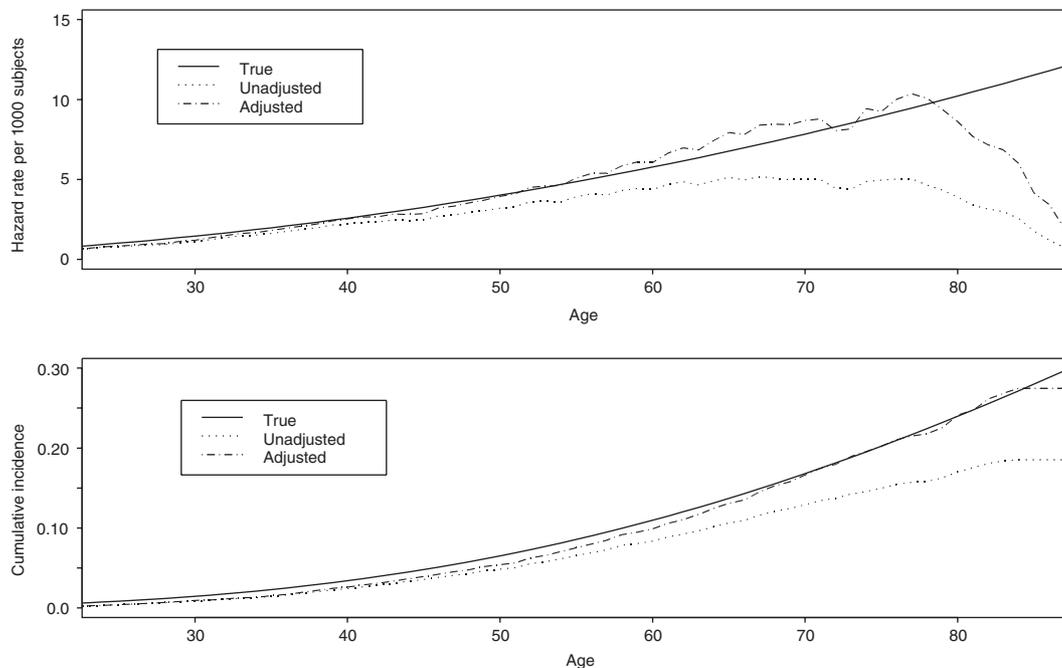


Fig. 2. Results from simulation experiments: Bias in estimation of age-specific hazard (above) and cumulative incidence (below) of ovarian cancer. Solid curves show hazard/cumulative-risk functions corresponding to true underlying Weibull distribution. Dotted lines show mean of estimates over 100 simulated data, when effect of mutation of risk of breast cancer is ignored. Dashed line shows corresponding mean estimates when effect of mutation on risk of breast cancer is accounted for, using method developed in this paper. Plots for hazard estimates are obtained after smoothing original estimates, using a moving average method that allows for 10-year window.

as a censoring/competing event for mortality and vice versa.

Figure 3 shows the estimate and the true values for the hazard and the cumulative incidence function of mortality in the carriers of the mutation. Clearly, ignoring the effect of the mutation on the competing risk of breast cancer causes noticeable bias, and after adjustment, the bias becomes much smaller. However, considering the fact that the competing event of breast cancer is strongly related to the mutation, the magnitude of the bias in the unadjusted estimate does not seem too great. This is likely related to the fact that in our simulation, (which is based on the real study setting), the most dramatic effect of BRCA1/2 mutations on the risk of breast cancer exists at younger ages ( $\leq 60$ ). The risk of mortality, on the other hand, is prominent at old ages ( $> 60$ ), and thus the estimation of risk of mortality, overall, is not very severely affected.

## CONCLUSIONS

One advantage of the cohort design is its ability to study the etiology of multiple diseases. In

certain settings, the kin-cohort design provides an attractive alternative to classical prospective cohort designs, as follow-up data on the cohort members of this design (i.e., the relatives) are collected very rapidly in a retrospective fashion through volunteers. For standard cohort data, it is well-known that for studying the etiology of a disease, one can ignore the censoring mechanism by assuming that the risks of the disease and censoring events are conditionally independent, given the exposures of interest. For a kin-cohort design, however, the censoring mechanism cannot be ignored, even if the independent censoring assumption holds, conditional on the mutation status of the relatives.

In kin-cohort analysis, estimation and interpretation of parameters while studying the effect of a gene depend on proper accounting for any other competing events that may be strongly influenced by the same gene. In general, the methods developed in this paper can be used to estimate the cause-specific hazards for different types of events without any assumption on the correlation between competing events. Based on these estimates, one can also compute other measures of

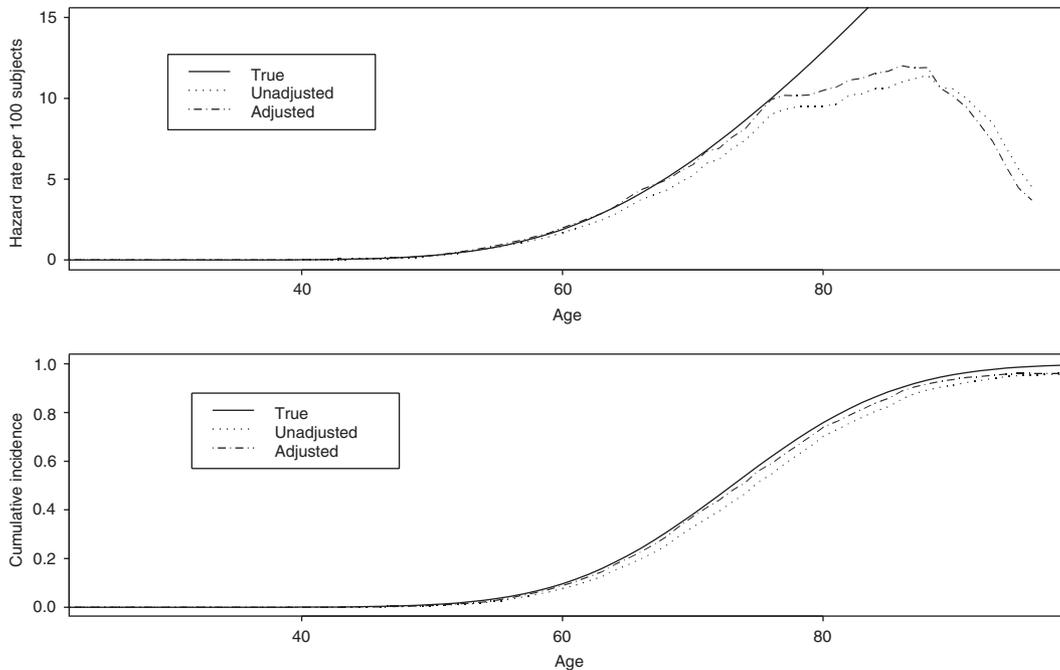


Fig. 3. Results from simulation experiments Bias in estimation of age-specific hazard (above) and cumulative incidence (below) of mortality. Solid curves show hazard/cumulative-risk functions corresponding to true underlying normal distribution. Dotted line shows mean of estimates over 100 simulated data, when effect of mutation on risk of breast cancer is ignored. Dashed line shows corresponding mean estimates when effect of mutation on risk of breast cancer is accounted for, using method developed in this paper. Plots for hazard estimates are obtained after smoothing original estimates, using a moving average method that allows for 5-year window.

risks, such as the cumulative incidence functions of the individual events in the presence of other events [Prentice et al., 1978; Gail et al., 1989], that are popularly used in the competing risk literature. Finally, although in this paper we focus on the kin-cohort setting, the proposed methodology can be used more generally for other types of cohort studies, where genotype information for some cohort members may be missing by design or by happenstance.

## REFERENCES

- Chatterjee N, Wacholder S. 2001. A marginal likelihood approach for estimating penetrance from kin-cohort designs. *Biometrics* 57:245–52.
- Efron B, Tibshirani RJ. 1993. An introduction to the bootstrap. London: Chapman and Hall.
- Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, Mulvihill JJ. 1989. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Nat Cancer Inst* 81:1879–1886.
- Gail MH, Pee D, Benichou J, Carroll R. 1999a. Designing studies to estimate the penetrance of an identified autosomal mutation: cohort, case-control, and genotyped-proband design. *Genet Epidemiol* 16:15–39.
- Gail MH, Pee D, Carroll R. 1999b. Kin-cohort designs for gene characterization. *J Nat Cancer Inst Monogr* 26:55–60.
- Godambe VP. 1991. Estimating functions. Oxford: Oxford University Press.
- Kalbfleisch JD, Prentice RL. 1980. The statistical analysis of failure time data. New York: John Wiley and Sons.
- Lee JS, Wacholder S, Struwing JP, McAdams M, Pee D, Brody LC, Tucker MA, Hartge P. 1999. Survival after breast cancer in Ashkenazi Jewish BRCA1 and BRCA2 mutation carriers. *J Nat Cancer Inst* 91:259–263.
- Liang KY, Zeger SL, Quaqish B. 1992. Multivariate regression analysis for categorical data. *J R Stat Soc Ser B* 54:3–40.
- Moore DF, Chatterjee N, Pee D, Gail MH. 2001. Pseudo-likelihood estimates of the cumulative risk of an autosomal dominant disease from a kin-cohort study. *Genet Epidemiol* 20:210–227.
- Prentice RL, Kalbfleisch JD, Peterson AV, Flournoy N, Farewell VT, Breslow NE. 1978. The analysis of failure times in the presence of competing risks. *Biometrics* 34:541–554.
- Rosen O, Jiang WX, Tanner MA. 2000. Mixtures of marginal models. *Biometrika* 87:391–404.
- Struwing JP, Hartge P, Wacholder S, Baker SM, Berlin M, McAdams M, Timmerman MM, Lawrence BC, Tucker MA. 1997. The risk of cancer associated with specific mutations of BRCA1 and BRCA2 among Ashkenazi Jews. *N Engl J Med* 336: 1401–1408.
- Wacholder S, Hartge P, Struwing JP, Pee D, McAdams M, Lawrence BC, Tucker MA. 1998. The kin-cohort study for estimating penetrance. *Am J Epidemiol* 148:623–630.

## APPENDIX

Here we derive Equation (4). We will use  $i = 1$  as an example. By Bayes' rule of conditional

probability, we can write

$$\begin{aligned}
 r_{1g}(t) &= \lim_{\delta t \downarrow 0} \frac{1}{\delta t} \Pr\{T_1 \in [t, t + \delta t) | T_1 \geq t, T_2 \geq t, G_0 = g\} \\
 &= \lim_{\delta t \downarrow 0} \frac{1}{\delta t} \frac{\Pr\{T_1 \in [t, t + \delta t), T_2 \geq t | G_0 = g\}}{\Pr\{T_1 \geq t, T_2 \geq t | G_0 = g\}} \\
 &= \lim_{\delta t \downarrow 0} \frac{1}{\delta t} \frac{\sum_{g'} \Pr\{T_1 \in [t, t + \delta t), T_2 \geq t | G = g'\} \Pr(G = g' | G_0 = g)}{\sum_{g''} \Pr\{T_1 \geq t, T_2 \geq t | G = g''\} \Pr(G = g'' | G_0 = g)} \\
 &= \sum_{g'} \left[ \lim_{\delta t \downarrow 0} \frac{1}{\delta t} \frac{\Pr\{T_1 \in [t, t + \delta t), T_2 \geq t | G = g'\}}{\Pr\{T_1 \geq t, T_2 \geq t | G = g'\}} \right] \\
 &\quad \times \frac{\Pr\{T_1 \geq t, T_2 \geq t | G = g'\} \Pr(G = g' | G_0 = g)}{\sum_{g''} \Pr\{T_1 \geq t, T_2 \geq t | G = g''\} \Pr(G = g'' | G_0 = g)}.
 \end{aligned}$$

In the third step of the above calculations, we implicitly assumed  $\Pr(T_1, T_2 | G = g', G_0 = g) = \Pr(T_1, T_2 | G = g')$ , i.e., given the relative's own genotype, the joint risk of two events in the relative does not depend on the genotype of the

proband. The proof of Equation (4) now follows by noting that in the product expression in the last line of above equations, the first term is  $\lambda_{1g}(t)$  (see Equation 2) and the second term is  $\Pr(G = g' | T_1 \geq t, T_2 \geq t, G_0 = g)$  (by Bayes' rule).