

# Association and Aggregation Analysis Using Kin-Cohort Designs With Applications to Genotype and Family History Data From the Washington Ashkenazi Study

Nilanjan Chatterjee,<sup>1\*</sup> Joanna Shih,<sup>2</sup> Patricia Hartge,<sup>1</sup> Lawrence Brody,<sup>3</sup> Margaret Tucker,<sup>1</sup> and Sholom Wacholder<sup>1</sup>

<sup>1</sup>*Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland*

<sup>2</sup>*Office of Biostatistics Research, National Heart, Lung and Blood Institute, Rockville, Maryland*

<sup>3</sup>*Laboratory of Gene Transfer, National Human Genome Research Institute, Bethesda, Maryland*

When a rare inherited mutation in a disease gene, such as *BRCA1*, is found through extensive study of high-risk families, it is critical to estimate not only age-specific penetrance of the disease associated with the mutation, but also the residual effect of family history once the mutation is taken into account. The kin-cohort design, a cross-sectional survey of a suitable population that collects DNA and family history data, provides an efficient alternative to cohort or case-control designs for estimating age-specific penetrance in a population not selected because of high familial risk. In this report, we develop a method for analyzing kin-cohort data that simultaneously estimate the age-specific cumulative risk of the disease among the carriers and non-carriers of the mutations and the gene-adjusted residual familial aggregation or correlation of the disease. We employ a semiparametric modeling approach, where the marginal cumulative risks corresponding to the carriers and non-carriers are treated non-parametrically and the residual familial aggregation is described parametrically by a class of bivariate failure time models known as copula models. A simple and robust two-stage method is developed for estimation. We apply the method to data from the Wash-

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

Received 20 July 2000; Accepted 2 November 2000

ington Ashkenazi Study [Struewing et al., 1997, *N Engl J Med* 336:1401–1408] to study the residual effect of family history on the risk of breast cancer among non-carriers and carriers of specific *BRCA1/BRCA2* germline mutations. We find that positive history of a single first-degree relative significantly increases risk of the non-carriers (RR = 2.0, 95% CI = 1.6–2.6) but has little or no effect on the carriers. *Genet. Epidemiol.* 21:123–138, 2001. © 2001 Wiley-Liss, Inc.

**Key words:** copula models; marginal likelihood; semiparametric estimation; residual familial aggregation; penetrance

## INTRODUCTION

An important goal in the study of the epidemiology of the familial diseases is to identify risk factors, genetic and/or environmental, which can explain familial aggregation of the disease. Once such a risk factor or a group of such factors is identified, investigators interested in public health and genetic counseling find it important to estimate not only the disease risk associated with the identified risk factors, but also the residual familial correlation of the disease after accounting for these known risk factors. Estimate of the residual familial correlation can be useful to guide further investigations of still unidentified risk factors and to build models for more accurate risk prediction. Breast cancer, for example, has been long known to aggregate within families. Published results indicate that in the general population approximately 10–15% of breast cancer cases have a family history of the disease and about 50% of these cases can be attributed to the inheritance of a susceptibility gene. Since the discovery of the *BRCA1/BRCA2* high-risk susceptibility genes, the quantification of the contribution of mutations in these genes has been a major research question. In high-risk families containing multiple cases of breast and/or ovarian cancer, mutations in these two genes account for the majority of the breast cancer cases [Easton et al., 1993, 1995]. In the general population, however, these mutations can explain only a small portion of the familial aggregation of the disease because of low prevalence, so other familial risk factors are suspected to play a role.

Struewing et al. [1997] and Wacholder et al. [1998] recently proposed and used a design known as kin-cohort for estimation of cancer risk associated with specific inherited mutations in disease genes. Struewing et al. [1997] genotyped 5,318 Ashkenazi Jewish volunteers living in the Washington, D.C., area for specific founder mutations in *BRCA1* and *BRCA2* genes. The volunteers also provided detailed personal and family history information, including age at onset, for a number of common cancers. From this study, using the disease history data on the relatives and the genotype information on the volunteers, the authors estimated the lifetime risk of penetrance of breast cancer, defined as the cumulative risk for a woman up to age 70, associated with the three Ashkenazi founder mutations, 185delAG and 5382insC in *BRCA1* and 617delT in *BRCA2*, to be 56%. Wacholder et al. [1998] termed the design underlying the study kin-cohort to emphasize the fact that the relatives of the volunteers formed a retrospective cohort who are followed from birth to onset of cancer or to the censoring time. The volunteers themselves, in contrast, cannot be treated as a cohort since a diseased individual could only participate if she remained alive until the study took place. Further complications arise since the mutations under study could also have an effect on the survival of the diseased individuals. Due

to the complex nature of the volunteers' disease data, the above authors did not use the personal history information of the volunteers in their analysis. Gail et al. [1999b] listed several practical advantages of the kin-cohort design over traditional cohort or case-control designs.

In this article, we proposed an approach for analyzing kin-cohort data that simultaneously addresses the problems of estimation of disease risk associated with the mutations under study and the estimation of the residual familial aggregation of the disease. Assuming that the relatives of the participants (volunteers in WAS study) in a kin-cohort study form a cohort, we consider modeling the joint age-at-onset distribution for two relatives given their genotypes using a class of bivariate failure time distributions, known as copula models. We treat the age-specific marginal cumulative risks among the non-carriers and carriers non-parametrically, while parametric models induced by the copula models are specified for the residual familial aggregation. Non-parametric estimation of the marginal cumulative risks not only allows robust estimation of the marginal penetrance functions, but also lets one investigate the residual correlation after accounting for the effect of the mutations to the fullest extent. On the other hand, non-parametric estimation of correlation among failure times in the presence of bivariate censoring is often complex and requires a large number of joint incidence of the disease. Given that joint incidence of breast cancer in two relatives is a rare event, for our application it seems best to estimate correlation using a summary measure induced by a parametric correlation model. For parameter estimation, we consider a two-stage quasi-likelihood estimation approach that takes account of the fact that the genotypes of the relatives are unknown, though indirect information is available through the genotyped participants. The method is computationally simple and requires minimal assumptions for consistent estimation of the parameters of interest. We apply the proposed method to the data from the Washington Ashkenazi Study to investigate the residual effect of family history on the risk of breast cancer among non-carriers and carriers of specific *BRCA1/BRCA2* mutations. This article is concluded with a discussion on the statistical methodologies and implications of the scientific findings.

## MODELS

### Copula Models

Let  $T_1$  and  $T_2$  denote two, possibly correlated, failure times. For example,  $T_1$  and  $T_2$  can be the age at onset of a disease for two related individuals. Suppose for  $i = 1, 2$ ,  $G_i(t_i) = \text{pr}(T_i \geq t_i)$  denotes the marginal survivor function for  $T_i$ . Let  $S(t_1, t_2) = \text{pr}(T_1 \geq t_1, T_2 \geq t_2)$  denote the joint survivor function. In the copula approach, one models  $S(t_1, t_2)$  in terms of the two marginal survivor functions,  $G_1$  and  $G_2$ , and a copula function which imposes a correlation structure between the two failure times. A copula function  $C(u, v)$  is a bivariate distribution function on the unit square  $[0, 1] \times [0, 1]$  with uniform marginal distributions. For any copula function  $C(u, v)$ , one can show that  $S(t_1, t_2) = C[G_1(t_1), G_2(t_2)]$  defines a joint survivor function, which has  $G_1$  and  $G_2$  as the corresponding marginal survivor functions. Several researchers in the past have studied various classes of copula functions, giving rise to different copula models. Three examples of such models are as follows:

(1) Frank's model [Frank, 1979]

$$C_{\theta}(u, v) = \begin{cases} \log \left\{ 1 - \frac{(1 - \theta^u)(1 - \theta^v)}{1 - \theta} \right\} / \log \theta, & 0 < \theta < 1 \\ uv, & \theta = 1 \end{cases}$$

(2) Clayton's model [Clayton, 1978]

$$C_{\theta}(u, v) = \begin{cases} (-1 + u^{-\theta} + v^{-\theta})^{-\frac{1}{\theta}}, & \theta > 0, \\ uv, & \theta = 0, \end{cases}$$

(3) Positive stable model [Hougard, 1986]

$$C_{\theta}(u, v) = \begin{cases} \exp \left[ -\{(-\log u)^{\frac{1}{\theta}} + (-\log v)^{\frac{1}{\theta}}\}^{\theta} \right] & 0 < \theta < 1 \\ uv, & \theta = 1, \end{cases}$$

Recent applications of these models in genetic epidemiology, particularly for the Clayton's model, include studies by Li et al. [1998] and Li and Huang [1998].

In the literature of failure time data analysis, both local and global measures of dependence are considered to measure association between two failure times. Local measures of association, such as the local odds ratio function [Oakes, 1989], are used to characterize correlation at specific time points of the sample space and can be used to investigate the time-dependent correlation structure between two failure times. On the other hand, global measures of association, such as Pearson's correlation coefficient or Kendall's tau, give a summary measure of association over the whole sample space. In all the three models described above,  $\theta$  can be interpreted as a summary measure of association, though the exact interpretation of  $\theta$  is different in different models. Many common global measures of association can be shown to have a one-to-one monotone relationship with  $\theta$  in the copula models. It is important to realize that even for a fixed global measure of association, different models can give quite different time-dependent association structures. The following illustrates this point. In analysis of familial disease data, epidemiologists often measure the familial correlation of the disease by the recurrence risk ratio: the ratio of the risk of the disease of a relative of an affected woman to that of a relative of an unaffected woman. Consider this recurrence risk ratio as a global measure of association between two failure times, where a woman is considered affected if she has the disease before a fixed old age, say 100 years. In Fig. 1, we illustrate through hypothetical examples how the different copula models imply different effects of the age at onset of the disease of a woman on the disease risk of her relative, even though the recurrence risk ratio is

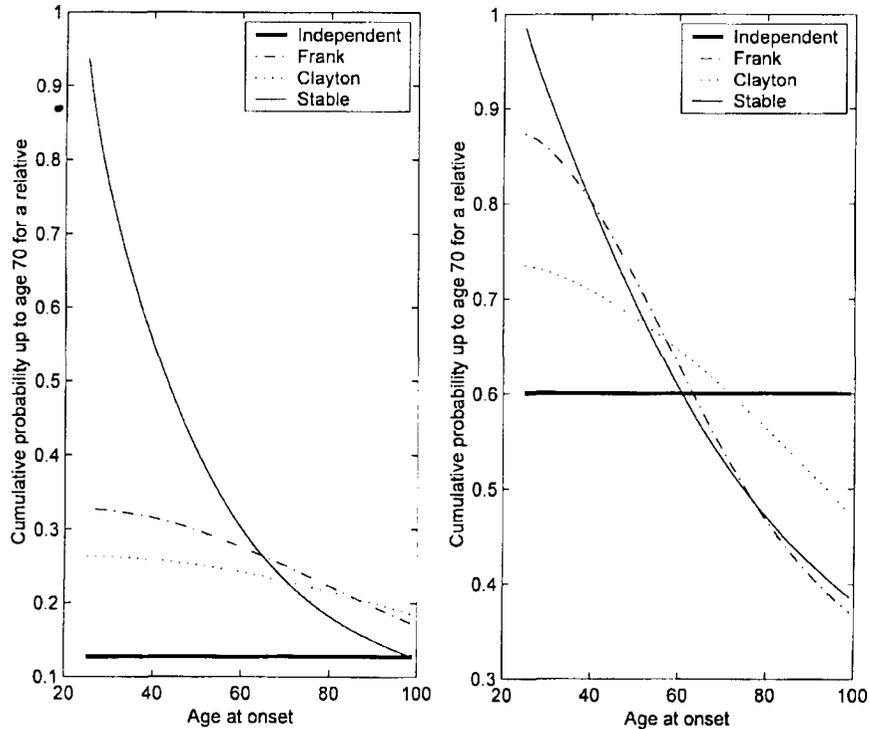


Fig. 1. The effect of age at onset of a disease of an individual (shown on  $x$  axis) on the disease risk of a relative (shown on  $y$  axis) as induced by different copula models. Plots are generated using hypothetical parameter values that are realistic for modeling breast cancer risk. **Left:** Corresponds to a population where the disease is rare (cumulative probabilities at age 50 and 70 are 0.013 and 0.045, respectively). **Right:** Corresponds to a population where the disease is more common (cumulative probabilities at age 50 and 70 are 0.33 and 0.56, respectively). The correlation parameters in the different models are chosen in such a way so that all the models correspond to a fixed recurrence risk ratio. The fixed values are 2.2 and 1.3 for left and right, respectively.

held constant. The plots are generated using hypothetical parameter values that are realistic for breast cancer data. In each plot, for a given value of the age at onset of an individual on the  $x$  axis, the  $y$  axis shows the cumulative risk until age 70 for a relative. It is assumed that there is no risk of disease below age 25. The joint distribution of the years to the disease after age 25 for two relatives is assumed to follow a copula model where the marginal distributions are the same for the two relatives. We considered a Weibull marginal distribution with two different sets of parameter values. The shape and scale parameter for the Weibull distribution are chosen to be 1.8595 and 0.0076 for the plot in Fig. 1 (left) and 1.4081 and 0.0209 for the plot in Fig. 1 (right), respectively. These parameter values are chosen so that Figure 1 (left) corresponds to a population where the disease is rare (cumulative probabilities at age 50 and 70 are 0.045 and 0.127), whereas Figure 1 (right) corresponds to a population where the disease is more common (cumulative probabilities at age 50 and 70 are 0.33 and

0.60). In each plot the association parameter ( $\theta$ ) corresponding to the three different copula models are chosen so that the recurrence risk ratio, as defined above, is 2.2 and 1.3 for the plot in Fig. 1 (left and right, respectively).

Figure 1 shows the different effect of the age at onset of a woman on her relative's risk to the disease corresponding to the different models. When the disease is rare (Fig. 1 left), for stable model we observe that early age at onset of a woman dramatically increases the risk of her relative. In contrast, early age at onset of a woman corresponds to only a slight and a moderate increase in risk of the relative in Clayton's and Frank's models, respectively. When the disease is more common (Fig. 1, right), young age at onset of a woman increases risk of her relative in all the models. Both Frank's and the stable models correspond to a steep increase, whereas Clayton's model corresponds to a more moderate increase.

#### APPLICATION OF COPULA MODEL TO WAS STUDY

We propose use of the copula functions to model the correlation between the ages at onset of pairs of female relatives of the participants, given their respective mutation status. Consider a participant who reported the personal history for each of his/her  $n$  relatives. Let  $g^p = 1$  or  $0$  indicate whether any of the mutant alleles are present or not in the participant. Similarly, let  $g_j^R, j = 1, \dots, n$  denote the mutation status of the relatives, which are unobserved in the kin-cohort design. Let  $T_j^R, j = 1, \dots, n$  denote the age at onset of the relatives, which may or may not be observed due to censoring. Let  $C_j^R, j = 1, \dots, n$  denote the censoring time for the relatives, which can be either their age at the time of the interview of the corresponding participant or their age at death, whichever occurred first. It is assumed that censoring time of the relatives are distributed independently of their genotypes and their ages at onset of the disease. Now let  $X_j^R = \min(T_j, C_j)$  denote the observed age for the  $j$ -th relative and  $\delta_j^R = 1$  or  $0$  denote the indicator of whether  $T_j \leq C_j$ , that is, whether the relative had the disease before the follow-up ended.

Let  $C_\theta(u, v)$  denote a class of copula functions, such that for any two relatives, say indexed by  $j$  and  $j'$ ,

$$\text{pr}(T_j \geq t_j, T_{j'} \geq t_{j'} | g_j^R, g_{j'}^R) = C_\theta[S_{g_j^R}(t_j), S_{g_{j'}^R}(t_{j'})] \quad (1)$$

for some "true" value of  $\theta = \theta_0$ , where  $S_g(t), g = 0, 1$  denote the survivor functions corresponding to non-carriers and carriers of the mutations, respectively. In equation (1),  $\theta$  gives a measure of association after accounting for the effect of the genotype status of the two relatives on their respective marginal risks. Thus, here  $\theta$  measures residual correlation of the disease, which cannot be explained by the gene under investigation. In the most general model, however,  $\theta$  still may depend on the mutation status of the members of the pairs as follows

$$\theta = \begin{cases} \theta_0, & \text{if both are non-carriers} \\ \theta_1, & \text{if exactly one is a carrier} \\ \theta_2, & \text{if both are carriers} \end{cases}$$

Unequal values of  $\theta$  are possible, for example, if the unobserved factors, genetic or environmental, that cause residual familial correlation have different effects on carriers and non-carriers. Thus, the difference in estimates of  $\theta_0$ ,  $\theta_1$ , and  $\theta_2$  will be suggestive of interaction between the mutations under study and the residual familial effect.

### Quasi-Likelihood

We construct a quasi-likelihood of the data by considering all possible pairs of relatives corresponding to a participant and treating the pairs as if they are independent of each other. The contribution of the  $(j, j')$  pair is given by the joint probability of their observed disease data given the genotype of the related participant. This probability can be obtained by first conditioning on the unknown genotypes of the relatives and then integrating with respect to the joint distribution of their genotypes given that of the participant as follows:

$$L_{j,j'} = \text{pr}(x_j^R, \delta_j^R, x_{j'}^R, \delta_{j'}^R | g^P) = \sum_{g_j^R, g_{j'}^R} \text{pr}(x_j^R, \delta_j^R, x_{j'}^R, \delta_{j'}^R | g_j^R, g_{j'}^R) \text{pr}(g_j^R, g_{j'}^R | g^P) \quad (2)$$

In equation (2), the first term inside  $\Sigma$  can be determined from our model after accounting for the possible single and double censoring (see Appendix), whereas the second term can be determined from the assumed mode of inheritance as a function of the allele frequency. The contribution of the relatives of a participant all together can be written as  $L = \prod_{(j, j') \in C} L_{j, j'}$ , where  $C$  denotes all possible pairs of relatives of the participant. The contribution of all the families together is obtained by taking the product of the contribution of the independent families. Though the quasi-likelihood ignores the dependency among the pairs within the same family, it can be shown that the parameter estimates obtained by maximizing this quasi-likelihood are consistent as long as the model for the bivariate distribution is correctly specified. A major motivation for considering one pair at a time is that we are only interested in assessing the correlation between two relatives at a time. Thus, higher order correlations can be treated as nuisance parameter and no assumptions need be made about them. Alternatively, one may model the joint distribution of all the relatives and consider a full likelihood approach. Although such an approach will be more powerful, consistency of the parameters of interest in this approach will depend on the correct specification of the whole multivariate model, which requires more model assumptions than the bivariate model. We also note that in our model the association parameter ( $\theta$ ) depends on the genotypes of the relatives. Although straightforward extensions of bivariate to multivariate copula models exist in the literature when  $\theta$  does not depend on covariates, such extensions can be complex when  $\theta$  depends on covariates.

### Two-Stage Estimation

If  $S_0(t)$  and  $S_1(t)$ , the age-specific marginal survival probability functions corresponding to non-carriers and carriers, are assumed to have known parametric forms with a small number of unknown parameters, the quasi-likelihood defined in the last section can be maximized jointly with respect to the parameters of the marginal distributions and the association parameters of the copula model. As the number of param-

eters needed to define  $S_0(t)$  and  $S_1(t)$  increases, however, the joint maximization problem becomes computationally challenging. In this article, we consider estimating the marginal survivor functions non-parametrically, which essentially involves defining  $S_0$  and  $S_1$  in terms of a “large” number of parameters. We proposed a two-stage estimation approach, in which at the first stage we estimate  $S_0$  and  $S_1$  using the marginal likelihood approach considered by Chatterjee and Wacholder [2001] and then estimate  $\theta$  using the quasi-likelihood defined above after replacing  $S_0$  and  $S_1$  by their respective estimate obtained at the first stage. In the marginal likelihood, the contribution of a family is found by taking the product of the marginal probabilities of the disease history of the individual relatives given the corresponding participant’s genotype. Thus, the contribution of a family corresponding to a participant with  $n$  relatives is given by  $\prod_{j=1}^n \text{pr}(x_j^R, \delta_j^R | g^P)$ , which can be written in terms of  $S_0$  and  $S_1$  as

$$\text{const} \times \prod_{j=1}^n \sum_{g_j^R=0}^1 \lambda_{g_j^R}^{\delta_j^R}(x_j^R) S_{g_j^R}(x_j^R) \text{pr}(g_j^R | g^P), \quad (3)$$

where  $\lambda_g(x)$  is the hazard function corresponding to  $S_g(x)$ . To obtain a non-parametric estimate of  $S_0$  and  $S_1$ , Chatterjee and Wacholder considered Kaplan-Meier product limit form for them, which allows a common set of jump points for the two curves—a jump at each age where at least one breast cancer onset has been observed among the pooled sample of the relatives of non-carrier and carrier participants. An EM algorithm is then defined to maximize the marginal likelihood to obtain estimates of the hazards corresponding to  $S_0$  and  $S_1$  at each of these jump points. We note that equation (3) does not depend on the choice of our copula model, and, thus, the corresponding estimates are free of the choice of the association model. In fact, it can be argued that the estimates of  $S_0$  and  $S_1$  will be consistent irrespective of the nature of the correlation between the members of a family, as long as the relatives of the participants can be treated as a random sample from an underlying population. In contrast, estimates of the marginal penetrance obtained from a full likelihood approach, the likelihood being formed under a particular assumption about the joint distribution of the relatives, may have considerable bias under misspecification of the joint model. Details of the marginal likelihood approach, together with its comparison with alternative approaches of penetrance estimation [Wacholder et al., 1998; Gail et al., 1999a] can be found in Chatterjee and Wacholder [2001]. Finally, the above two-stage estimation approach can be considered as an extension of the two-stage estimation approach considered by Shih [1998] for ordinary copula models to our setting where data are observed from mixtures of copula models.

The copula parameters, although they give a measure of residual association, are perhaps too abstract to be useful for interpretation of the results. More interpretable quantities can be easily computed from the estimate of the marginal penetrances and the copula parameters. For example, a popular measure of familial aggregation often used in genetic epidemiology is the recurrence risk, the risk of a woman given the history of the disease in one of her relatives. For measuring residual aggregation, we compute the recurrence risk for a woman given her mutation status using a formula given in the Appendix.

For variance estimation we consider a bootstrap approach, where the families are treated as the bootstrap sampling units to account for the correlation within a family. First, we sample with replacement from all the families, to obtain a random sample of the same size as the total number of families in the study. For each such sample, we first obtain the marginal-likelihood estimate of the marginal survivor functions and then substitute those in the quasi-likelihood for the corresponding population quantities. The quasi-likelihood is then maximized to obtain the estimate of the copula parameters for the corresponding bootstrap sample. Bootstrap variance estimates of parameters of interest can be computed as the variance of the corresponding estimates over a large number of bootstrap samples. Shih [1998] considered a two-stage quasi-likelihood approach for inference in bivariate copula models. Though her problem was somewhat simpler, the asymptotic theory she developed for copula models with parametric marginal distributions can be used in our setting to derive an asymptotic variance formula if we assume a parametric or discrete marginal hazard model for carriers and non-carriers. Asymptotic theory for non-parametric marginal hazards will require modern semi-parametric estimation theory and is beyond the scope of this paper.

## RESULTS FROM THE WASHINGTON ASHKENAZI STUDY

We applied the two-stage estimation approach to the data from the Washington Ashkenazi Study to estimate the residual effect of family history in the carriers and the non-carriers of *BRCA1/BRCA2* mutations. Estimation of the marginal cumulative risks at the first stage involves 305 first-degree relatives of 110 carrier volunteers and 12,991 first-degree relatives of 4,752 non-carrier volunteers. Since only female breast cancer is considered, only female first-degree relatives (mother, sister, or daughter) of the participants (male or female) are included in the analysis. Assuming Hardy-Weinberg equilibrium and treating the volunteers as a random sample, Struewing et al. [1997] estimated the allele frequency of the *BRCA1/BRCA2* mutations in the Ashkenazi population from the carrier frequency of the volunteers. We used this estimate of allele frequency (0.016) for our analysis of the data. The marginal likelihood estimate of the cumulative risk among the carriers of any of the three specific mutations and the corresponding estimate given in Struewing et al. [1997] are shown in Fig. 2. We observed that the new estimate followed the original estimate very closely until about age 60. After age 60, the original reported estimate was non-monotonic; a natural monotonicity constraint is automatically imposed in the marginal likelihood estimator described in the Two-Stage Estimation section. The new estimates of the cumulative risk up to ages 70 and 80 were 0.60 and 0.67, as compared to the original estimates of 0.56 and 0.60, respectively. The marginal-likelihood risk estimate for the non-carriers was almost identical to that reported by Struewing et al. (data not shown).

In the second stage, we considered all possible pairs of the first-degree relatives of the non-carrier and carrier volunteers, who were also first-degree relatives of each other. Thus, with respect to the relationship to the volunteers, three types of pairs were considered, namely mother-sister, sister-sister, and daughter-daughter. Table I shows the frequency distribution of these pairs by the disease status of the members of the pairs. As described in the Application of Copula Model to WAS Study section,

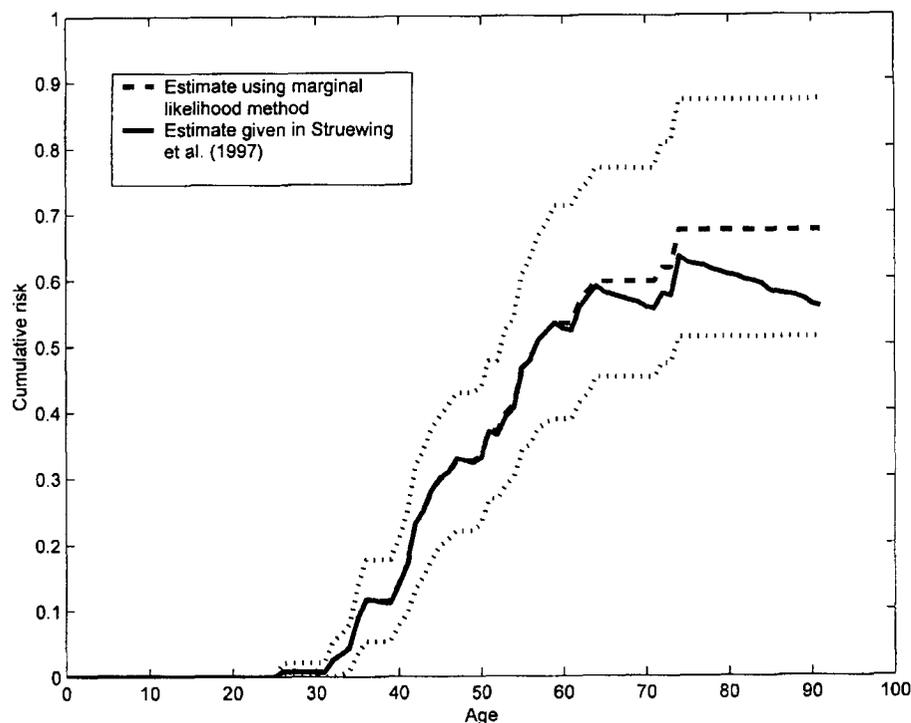


Fig. 2. The age-specific cumulative risk of breast cancer among carriers and non-carriers of any three *BRCA1/BRCA2* mutation: original estimates reported in Struwing et al. [1997] and new estimates based on the marginal likelihood method. The dotted line is the bootstrap 95% pointwise confidence interval for the marginal likelihood estimate.

the most general model for our data will be to allow three different copula parameters,  $\theta_0$ ,  $\theta_1$ , and  $\theta_2$ , which respectively measure the association between two non-carriers, one non-carrier and one carrier, and two carriers. We note that the *BRCA1/BRCA2* mutations are very rare, so the relatives are very unlikely to be carriers unless the related volunteer is a carrier; thus, most of the information on  $\theta_1$  and  $\theta_2$  comes from the pairs corresponding to the carrier volunteers. From Table I, we note that of 160 such pairs, only eight were composed of two affected members. Due to such sparseness of the data, we found that estimation of  $\theta_1$  and  $\theta_2$  separately is difficult in our data set. Although we could obtain these estimates by maximizing the

**TABLE I. Frequency Distribution of the Pairs of First Degree Relatives of Non-Carrier and Carrier Volunteers by Breast Cancer Status of the Members of the Pairs**

Relatives affected	Non-carriers	Carriers
0	5,759	106
1	838	46
2	76	8
Total	6,673	160

quasi-likelihood defined earlier, the resulting estimates were numerically unstable. As the next best approximation, we fitted the following two-parameter model

$$\theta = \begin{cases} \theta_0 & \text{if both non-carriers} \\ \theta_1^* & \text{at least one is a carrier} \end{cases}$$

Table II shows the genotype-specific recurrence risk, defined as the cumulative risk for a woman up to age 70 given her carrier status and the disease history of a 70-year old first-degree relative. Results from all three models indicate that there is a strong and statistically significant effect of family history on non-carriers, but the effect of family history on carriers is small or none. Specifically, in both Frank's and Clayton's models, the relative risk associated with the relative's disease status (recurrence risk ratio) among non-carriers is 2.1 (95% CI: 1.6–2.6), whereas the corre-

**TABLE II. Estimate of Genotype-Specific Recurrence Risk: Cumulative Risk of Breast Cancer for a Woman Until Age 70, Given Her Carrier Status and the Disease History of a 70-Year-Old Female First-Degree Relative\***

Relative's status	Relative's age at diagnosis	Non-carriers		Carriers	
		Estimate <sup>a</sup> (%)	95% CI <sup>b</sup>	Estimate <sup>a</sup> (%)	95% CI <sup>b</sup>
Frank's model: (log QL = -6,708.3)					
No Hx		11	10–12	57	41–77
Hx		23	19–28	63	48–81
	< 30	25	20–31	67	48–89
	30–39	24	20–30	66	48–87
	40–49	24	20–30	64	48–84
	50–59	23	19–28	62	48–79
	60–69	23	19–27	62	48–78
Clayton's model: (log QL = -6,708.8)					
No Hx		11	10–12	56	40–77
Hx		23	19–28	67	48–85
	< 30	24	20–30	70	48–93
	30–39	23	19–29	70	43–92
	49–49	23	19–29	68	46–88
	50–59	23	19–28	65	49–81
	60–69	22	19–27	65	48–80
Stable model: (log QL = -6,709.7)					
No Hx		12	11–13	60	44–78
Hx		20	17–26	60	46–78
	< 30	42	29–57	60	46–82
	30–39	29	22–38	60	46–80
	40–49	23	18–29	60	46–78
	50–59	19	16–23	60	46–78
	60–69	17	15–20	60	46–78

\*For women with a diseased relative, effect of the age at onset of the relative is also shown by computing the recurrence risk corresponding to five intervals for the age at diagnosis of the relative.

<sup>a</sup>Estimates are obtained using the copula model and related estimation methods described in Quasi-Likelihood and Two-Stage Estimation sections.

<sup>b</sup>Based on 150 bootstrap sample of the families.

No Hx, without disease; Hx, with disease.

sponding relative risk in the stable model is 1.8 (1.4–2.4). Among carriers, the corresponding relative risks for Frank's, Clayton's, and stable models are 1.1 (1.0–1.7), 1.2 (1.0–1.94), and 1.0 (1.0–1.3), respectively. The lower confidence bound of 1.0 for each of these relative risk estimates results from the fact that in our models we do not allow for negative association among family members and relative risk less than 1.0 is not possible. We also note that for the stable model the estimate of  $\theta^*$  was at the boundary of the parameter space, which corresponded to no association. Table II also shows the effect of age at onset of the relative on the recurrence risk estimate. For Clayton's and Frank's models, these estimates show very little dependence on the relative's age at onset. For both non-carriers and carriers, a slight increase in the risk is observed with younger age at onset of the relative. For the stable model, on the other hand, a relative with early age at onset seems to significantly increase the risk of a non-carrier woman. These age-dependent estimates, however, are mostly model driven, since, as discussed in the Copula Models section, the choice of a particular copula model specifies a particular time-dependent association structure. As we have seen in Fig. 1, compared to the Frank's and Clayton's models, the stable model specifies a very strong effect of time on the association between two relatives and, thus, the differences in the respective time-dependent estimates that we observe are not surprising. To see which type of association model best fits a particular set of data, a model selection criterion is necessary. Since all the three models were fitted using the same quasi-likelihood described in the Quasi-Likelihood section and all of them involve the same number of parameters, we used the value of the log(quasi-likelihood) at the estimated parameter values to compare the fit of the different models in our data. The smallest value of the log(quasi-likelihood) for Frank's model suggests this to be the best fit for our data. However, we notice that the differences among the fit of the different models are not very large.

## DISCUSSION

Our analysis of WAS data suggests that the presence of breast cancer in a single first-degree relative significantly increases the risk of non-carriers but has little or no effect on the risk of carriers. This conclusion about the global effect of family history, which considers the effect of the relative having the disease before age 70 without any regard to the actual age at onset, seems to be fairly robust to the choice of different association models. The estimates showing the effect of age at onset of the relative, however, are sensitive to the choice of the model and have to be interpreted cautiously. Although the best-fitting model suggests little effect of age at onset of relatives, its fit was only marginally better than an alternative model that assumes strong effect of age at onset on the risk of non-carriers. It seems that more data will be necessary to make a more definitive conclusion about the effect of age at onset.

Claus et al. [1998] and Kauffman and Struewing [1998] reported estimates of odds ratios for positive family history among non-carriers of *BRCA1/BRCA2* mutations. The outcome data for the first study consisted of the disease status of the cases and controls who participated in the Cancer and Steroid Hormone case-control study, while that of the second study consisted of the disease status of the WAS volunteers. Both studies treated the subjects' family history as an exposure in their analysis and estimated the corresponding odds ratio using standard logistic regression methods.

The overall conclusion that positive family history increases the risk of breast cancer among non-carriers is the same in these studies as ours. The effect of positive disease history of a single relative from the analysis of Kauffman and Struewing, however, was weaker, the estimated odds ratio for first-degree family history effect being about 1.5 (95% CI: 1.1–2.2), in contrast to those from Claus et al. and the current analysis, both of which estimated the corresponding odds ratio to be about 2.0.

Certain unique features and caveats of our analysis in comparison to the two other studies merit discussion. Since the outcome data for our analysis are obtained from a retrospective cohort formed by the first-degree relatives of the volunteers, we can estimate absolute risks as well as odd ratios. Second, we attempted to estimate the effect of family history on carriers. Third, our analysis and that of Claus et al. [1998] account for mutation status for individuals with no directly available genotype information in different ways. We account for the unknown mutation status more formally at an individual level, while Claus et al. [1998] used an ad hoc approach that predicts the carrier status at a group level. The statistical program that they used to predict the carrier probabilities involves external estimates of the penetrance functions associated with *BRCA1/BRCA2* mutation status. It is unclear whether and how to incorporate uncertainties in these estimates in their final analysis. Furthermore, to predict the carrier probabilities their method also assumes that residual correlation is absent, even though the primary goal of the study is investigation of residual family history effect. In contrast, our estimation method is able to avoid these issues by using internal data exclusively. Kauffman and Struewing, on the other hand, based their analysis on the female volunteers of the WAS study and formulated the problem in terms of estimating the risk of non-carrier volunteers by their family history. As noted by the authors and discussed before in this article, use of the volunteer's disease may produce biased results due to unknown influence of family history and the mutations under study on survival of the cases [see also Chappuis, 1999] and hence their ability to participate in the study. Due to the cohort nature of the relatives' disease incidence data, our analysis is not expected to be affected by the survival of the diseased relatives after their cancer incidence.

A limitation of the WAS study, as discussed by Struewing et al. [1997] and Wacholder et al. [1998], is reliance on volunteers. It is likely that volunteers with positive family history will be more likely to participate in the study, the chance of participation probably increasing with the number of affected relatives. Our analysis, which throughout assumes that the relatives can be treated as a cohort of individuals randomly selected from an underlying population, does not account for such volunteer bias. Our estimates of marginal and conditional risks, based on families that possibly have more affected relatives than families in the general population, can be expected to be slightly higher than the true risks in the general population.

The observed strong effect of family history on the risk of breast cancer of non-carrier females can be due to a combination of other breast cancer susceptibility genes and environmental factors that tend to aggregate within a family. The environmental factors may include some known risk factors, such as age at first birth and age at menarche, for both of which correlation has been reported among related individuals [Treloar and Martin, 1990], and unknown risk factors, with the most likely candidates being dietary and lifestyle factors. Simulation studies [Khoury et al., 1988], however, show that aggregation of a single environmental factor can have a signifi-

cant effect on aggregation of a disease only if the relative risk of the disease corresponding to the factor is extremely high. On the other hand, there has been some evidence for existence of at least one major susceptibility gene in addition to *BRCA1/BRCA2* [Serova et al., 1997]. Research to identify such genes is in progress. A number of candidate breast cancer susceptibility genes, including but not limited to *p53* and *ATM*, are being studied, although the importance of their contribution to the risk of breast cancer is not yet well understood. The observed small effect of family history among the *BRCA1/BRCA2* carriers in these data indicates that the combined effect of genetic and environmental factors, which may explain the family history effect in non-carriers, is weak on carriers. This is consistent with other studies that have found that various known risk factors for breast and ovarian cancer in the general population have little, none, or even opposite effects on the risk of *BRCA1/BRCA2* carriers. Jernstorm et al. [1999], for example reported that early age at first birth, which is known to be associated with reduced risk of breast cancer in the general population, did not have any effect on the risk of carriers. The same study also found that increasing parity, which is believed to be protective for breast cancer, was associated with increasing risk of breast cancer among *BRCA1/BRCA2* carriers.

In summary, we have proposed statistical methods for analyzing data from kin-cohort designs to estimate the risk of a disease associated with certain known mutations and simultaneously measure the residual familial aggregation of the disease after accounting for these mutations. The basic modeling approach consists of specifying the joint age-at-onset distribution for pairs of relatives conditional on their genotypes using a class of models for bivariate failure time data, known as copula models. The estimation method we proposed is simple to implement, computationally fast even when the marginal penetrances are treated non-parametrically, and does not require specification of third or higher order associations among three or more relatives. Applying the methods to the WAS data, we obtained non-parametric estimate of the age-specific cumulative risk of breast cancer among the carriers and the non-carriers of *BRCA1/BRCA2* mutations, which corrected the non-monotonicity problem in the corresponding original estimates reported by Struewing et al. [1997], and also estimated the residual effect of family history, for both non-carriers and carriers. In this paper, we concentrated on parametric models for residual correlation. An important area of future research will be to consider more non-parametric estimation of the age-dependent correlation structure, particularly among the non-carriers for whom the presence of a residual familial effect seems to be quite evident. A particular pattern of such non-parametric estimates, such as stronger correlation at young ages or uniform correlation over all ages, may shed light on mechanisms of action for the residual familial risk factors.

## ACKNOWLEDGMENTS

We thank Dr. Mitchell Gail for his insightful comments on this paper that improved the presentation.

## REFERENCES

- Chappuis PO. 1999. The influence of familial and hereditary factors on the prognosis of breast cancer. *Ann Oncol* 10:1163–70.

- Chatterjee N, Wacholder S. 2001. A marginal likelihood approach for estimating penetrance from kin-cohort designs. *Biometrics* 57:245–52.
- Claus EB, Risch NJ, Thompson WD. 1990. Age at onset as an indicator of familial risk of breast cancer. *Am J Epidemiol* 131:961–72.
- Claus EB, Schildkraut ES, Edwin IS Jr, Berry D, Parmigiani G. 1998. Effect of BRCA1 and BRCA2 on the association between breast cancer risk and family history. *J Natl Cancer Inst* 90:1824–9.
- Clayton DG. 1978. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* 65:141–51.
- Easton DF, Bishop DT, Ford D, Crockford GP. 1993. Genetic linkage analysis in familial breast and ovarian cancer: results from 214 families. The breast cancer linkage consortium. *Am J Hum Genet* 52:678–701.
- Easton DF, Ford D, Bishop DT. 1995. Breast and ovarian cancer incidence in BRCA1-mutation carriers. The breast cancer linkage consortium. *Am J Hum Genet* 56:265–71.
- Frank MJ. 1979. On the simultaneous associativity of  $F(x,y)$  and  $x + y - F(x,y)$ . *Aequationes Mathematicae* 19:194–226.
- Gail M, Pee D, Benecrou 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 M, Pee D, Carroll R. 1999b. Kin-cohort designs for gene characterization. *J Natl Cancer Inst Monogr* 26:55–60.
- Hougaard P. 1986. A class of multivariate failure time distributions. *Biometrika* 73:671–8.
- Jemstrom H, Lerman C, Ghadirian P, Lynch HT, Weber B, Daly M, Olopade OI, Foulkes WD, Warner E, Brunet JS, Narod SA. 1999. Pregnancy and risk of early breast cancer on carriers of BRCA1 and BRCA2. *Lancet* 354:1846–50.
- Kauffman DJ, Struwing J. 1999. Re: Effect of BRCA1 and BRCA2 on the association between breast cancer risk and family history. *J Natl Cancer Inst* 91:1250.
- Khoury MJ, Beaty TH, Liang KY. 1988. Can familial aggregation of disease be explained by familial aggregation of environmental risk factors? *Am J Epidemiol* 127:674–83.
- Li H, Huang J. 1998. Semiparametric linkage analysis using pseudolikelihoods on neighbouring sets. *Ann Hum Genet* 62:323–36.
- Li H, Yang P, Schwartz AG. 1998. Analysis of age at onset data from case-control family studies. *Biometrics* 54:1030–9.
- Oakes D. 1989. Bivariate survival models induced by frailties. *J Am Stat Assoc* 84:487–93.
- Serova OM, Mazoyer S, Puget N, Dubois V, Tonin P, Shugar Y, Goldgar D, Narod SA, Lynch HT, Lenoir GM. 1997. Mutations in BRCA1 and BRCA2 in breast cancer families: are there more breast cancer susceptibility genes? *Am J Human Genet* 60:486–95.
- Shih JH. 1998. Modeling multivariate discrete failure time data. *Biometrics* 54:1115–28.
- 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* 336:1401–8.
- Treloar SA, Martin NG. 1990. Age at menarche as a fitness trait: nonadditive genetic variance detected in a large twin sample. *Am J Hum Genet* 47:137–48.
- 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* 148:623–30.

## APPENDIX

### Details for Computing Quasi-Likelihood

In equation (2),  $\text{pr}(t_j^R, \delta_j^R, t_j^R, \delta_j^R | g_j^R, g_j^R)$  can have three different forms corresponding to pairs with none, exactly one, or two diseased members. In most applications, including the current one, ages are recorded up to the nearest integer. In that case, the form of  $\text{pr}(t_j^R, \delta_j^R, t_j^R, \delta_j^R | g_j^R, g_j^R)$ , after ignoring constant terms that do not depend on the parameters of the model, are as follows: for  $\delta_j^R = 0$  and  $\delta_j^R = 0$ ,

$$\text{pr}(T_j^* \geq t_j^R, T_{j'}^* \geq t_{j'}^R | g_j^R, g_{j'}^R) = C_\theta[S_{g_j^R}(t_j^R), S_{g_{j'}^R}(t_{j'}^R)],$$

for  $\delta_j^R = 1$  and  $\delta_{j'}^R = 0$ ,

$$\text{pr}(T_j^* = t_j^R, T_{j'}^* \geq t_{j'}^R | g_j^R, g_{j'}^R) = C_\theta[S_{g_j^R}(t_j^R - 1), S_{g_{j'}^R}(t_{j'}^R)] - C_\theta[S_{g_j^R}(t_j^R), S_{g_{j'}^R}(t_{j'}^R)],$$

for  $\delta_j = 1$  and  $\delta_{j'} = 1$ ,

$$\begin{aligned} & \text{pr}(T_j^* = t_j^R, T_{j'}^* = t_{j'}^R | g_j^R, g_{j'}^R) \\ &= C_\theta[S_{g_j^R}(t_j^R - 1), S_{g_{j'}^R}(t_{j'}^R - 1)] - C_\theta[S_{g_j^R}(t_j^R), S_{g_{j'}^R}(t_{j'}^R - 1)] \\ & \quad - C_\theta[S_{g_j^R}(t_j^R - 1), S_{g_{j'}^R}(t_{j'}^R)] + C_\theta[S_{g_j^R}(t_j^R), S_{g_{j'}^R}(t_{j'}^R)] \end{aligned}$$

### Computing Genotype Specific Recurrence Risk

The recurrence risk defined as the cumulative risk for a woman up to a specific age, say  $t$ , given her mutation status ( $g_0$ ) and the age at onset of the relative ( $T_1$ ), can be computed by the formula

$$\text{Pr}(T_0 \leq t | a \leq T_1 \leq b, g_0) = \frac{\sum_{g_1} \text{Pr}(T_0 \leq t, a \leq T_1 \leq b | g_0, g_1) \text{Pr}(g_1 | g_0)}{\sum_{g_1} \text{Pr}(a \leq T_1 \leq b | g_1) \text{Pr}(g_1 | g_0)} \quad (4)$$

where  $T_0$  denotes the age at onset of the woman,  $(a, b]$  denotes an age interval containing the relative's age at onset, and  $g_1$  denotes the unknown genotype of the relative. If the relative was alive without the disease until a specific age, say  $s$ , one can choose  $a = s$  and  $b = \infty$  in the above formula. In equation (4),  $\text{Pr}(a \leq T_1 \leq b | g_1)$  can be estimated from the estimates of the marginal risks,  $S_0(t)$  and  $S_1(t)$ , and  $\text{Pr}(T_0 \leq t, a \leq T_1 \leq b | g_1, g_0)$  can be estimated from the estimates of  $S_0(t), S_1(t)$  and  $\theta$  using the model formula given in (1).