Journal of Biometrics & Biostatistics
ISSN: 2155-6180
Choi, J Biomet Biostat 2012, S1
DOI: 10.4172/2155-6180.S1-006
Research Article
Open Access

A Frailty-Model-Based Method for Estimating Age-Dependent Penetrance from Family Data

Yun-Hee Choi*

Department of Epidemiology and Biostatistics, Western University, London, ON, Canada

Abstract

Accurate estimates of disease risk (penetrance) associated with inherited gene mutations are critical for the clinical management of individuals at risk, but this estimation raises many statistical challenges especially when performed in a family-based design. In this paper, we propose a general frailty model-based approach to accommodate this design, where the frailty random effect accounts for shared risk among family members not due to the observed risk factors. It is of major interest when the goal is to discover other genetic variations besides the major gene and to get accurate estimates of penetrance (i.e. unbiased by unknown confounding factors). This approach is further extended to accommodate missing genotypes in family members and the non-random ascertainment of the families. Simulation results show that the proposed method performs well in realistic settings. Finally, a family-based breast cancer study of the BRCA1 and BRCA2 genes is used to illustrate the method.

Keywords: Frailty; Risk estimation; Correlated survival times; Ascertainment; Missing data; Gene mutation

Introduction

In genetic epidemiology, family data are often used to study genetically transmitted diseases as families sampled from affected individuals tend to include more cases and thus are more likely to harbor a disease gene mutation or a gene variant. Once a major gene related to a disease has been identified, the next step consists in trying to characterize and estimate its associated risks in the population (relative and absolute risks, allele frequency, attributable fraction, etc.). This could have important scientific and public health implications for developing intervention and prevention strategies for those genetically susceptible individuals. With the identified major genetic factors, the estimation of additional residual familial correlation is also needed to make correct inference about the major gene effect and to identify additional residual factors that could also have a genetic origin. The major gene effect associated with a disease can be expressed as the lifetime risk in gene carriers i.e., age-specific penetrance as a probability of developing a particular disease among mutation carriers when onset varies with age and relative risk between carriers and non-carriers of the mutated gene. Recent studies [1,2,3] have developed ascertainment-corrected likelihood approaches for estimating such disease risks under various family-based study designs and for different genetic models. They have provided nearly unbiased estimates of penetrance except in the presence of a second gene effect or other causes of residual familial correlations. To directly model the residual familial correlations induced by unknown risk factors, we propose a frailty-based likelihood approach. In this paper, a shared frailty model is incorporated to better characterize the disease risks associated with identified gene mutations and familial correlation by taking the sampling design and ascertainment correction into account. The term frailty was first introduced by Vaupel et al. [4] in the survival model framework to account for unobserved heterogeneity in the study population. The use of the frailty model to describe population heterogeneity was also studied by several authors [5,6,7]. Further, the frailty model was extended to accommodate unknown common risk factors such as common environmental or genetic factors shared within clusters [8,9,10]. More recently, various authors utilized a family-specific frailty to describe the dependence within a family [11-16]. This model allows each family to share a random variable, which refers to as a frailty, to explain an unknown common risk factor within families. On the other hand, the independent model is occasionally used by simply ignoring the frailty in the model but with adjustment of familial correlation using a robust variance estimator. Keiding et al. [17] showed that ignoring the frailty effect can lead to an underestimation of the covariate effects under the misspecified model. Gong and Whittemore [18] found that the presence of additional risk factors, such as a second gene, could result in an upward bias in risk estimates.

The main objective of this paper is to develop a general framework for the shared frailty model for analyzing correlated time-to-event data arising from a family-based study. Analysis of this design also requires an ascertainment correction to make appropriate inference and inference about missing genotype data.

This paper begins with a description of a shared frailty model, followed by the likelihood formulations for the family-specific frailty model and the independent model (Section 2). These approaches are extended to account for the non-random ascertainment of family data. We then introduce robust variance estimators for the log relative risk and the cumulative lifetime risk (penetrance) associated with a major gene of interest. In Section 3, a simulation study examines the performance of the frailty and independent models in terms of bias and precision in the estimation of the model parameters. Section 4 illustrates an application of the frailty model to an early-onset breast cancer study under complex sampling design from three population-based family registries. Concluding remarks follow in Section 5.

***Corresponding author:** Yun-Hee Choi, Department of Epidemiology and Biostatistics, Western University, London, ON, Canada, Tel: 519-661-2111 (86256); Fax: 519-661-3766; E-mail: Yun-Hee.Choi@schulich.uwo.ca

Received September 08, 2011; **Accepted** February 11, 2012; **Published** February 15, 2012

Citation: Choi YH (2012) A Frailty-Model-Based Method for Estimating Age-Dependent Penetrance from Family Data. J Biomet Biostat S1:006. doi:10.4172/2155-6180.S1-006

Methods

A frailty model for family-based studies

Consider that the data arise from n families and each family consists of n_f individuals, where f indexes families, f =1,…,n, and i indexes individuals, i =1,…,n_f.

The following is the shared frailty model for the time to onset T_{fi} for individual i from family f. We note that this shared frailty model is more appropriate for nuclear families. Let Z_f denote the frailty shared within family f. Given frailty $Z_f = z_f$ and a vector of observed covariates X_i for individual i, the conditional proportional hazard model and the corresponding survivor function are given by

$$h_{fi}(t_{fi} \mid z_f, X_i) = z_f h_0(t_{fi}) e^{X_i\beta};$$

$$S_{fi}(t_{fi} \mid z_f, X_i) = \exp\{-z_f H_0(t_{fi}) e^{X_i\beta}\}, \quad (1)$$

where $h_0(.)$ and $H_0(.)$ represent the baseline hazard and cumulative hazard functions, respectively.

The frailty model assumes that, conditional on the value of the frailty, Z, and the observed covariates, X, which include the mutation status and other covariates, the observations in a family are independent and the association between family members is a consequence of the frailty distribution whose density function is $g(.; k)$, depending on the frailty parameter k. Thus, the frailty parameter, k, determines the dependence within families. For mathematical simplicity, we use gamma frailty [11,15,19] with mean 1 and variance $1/k$ since it has a close-form expression for the penetrance and likelihood functions. The frailty parameter, k, measures the magnitude of the dependence among ages of onset from family members, with a smaller value of k implying a stronger dependence. For the baseline hazard function, we use a Weibull distribution $h_0(t) = \lambda\rho(\lambda t)^{\rho-1}$ as it enables flexible modeling of the baseline hazard, in particular with constant, increasing or decreasing hazards.

Based on this model, our interest is in the following disease risks: relative risk and penetrance.

Relative risk: If we consider that the model includes a genetic variable that indicates the mutation carrier status, then the corresponding β represents the log hazards ratio between carriers and non-carriers of a mutated gene. Then, exp(β) is referred as to relative risks of disease between carriers and non-carriers.

Penetrance: The penetrance is the age-specific probability of developing a disease by a certain age t. It can be formulated as a cumulative distribution function of t given observed covariates X, i.e., $F(t|X)$, which can be derived from equation (1) by integrating over the unknown frailty. Then, we can express the penetrance at age t as

$$F(t \mid X) = 1 - \int_{z_f} S(t \mid z_f, X) g(z_f) dz_f$$

$$= 1 - \{1 + (\lambda t)^{\rho} e^{X\beta} / k\}^{-k}, \quad (2)$$

using the Laplace transform of the gamma frailty distribution; details follow below.

Likelihood construction

Based on the conditional model given the frailty and observed covariates described in (1), we obtain the likelihood function for family f by integrating over the frailty distribution as

$$L_f(\theta) = \int \prod_{i=1}^{n_f} h(t_{fi} \mid z_f)^{\delta_{fi}} S(t_{fi} \mid z_f) g(z_f) dz_f \quad (3)$$

$$= \int \prod_{i=1}^{n_f} \{h_0(t_{fi}) z_f e^{X_i\beta}\}^{\delta_{fi}} e^{-z_f H_0(t_{fi}) e^{X_i\beta}} g(z_f) dz_f$$

$$= \prod_{i=1}^{n_f} \{h_0(t_{fi}) e^{X_i\beta}\}^{\delta_{fi}} \times \int z_f^{d_f} e^{-z_f \sum_i H_0(t_{fi}) e^{X_i\beta}} g(z_f) dz_f$$

$$= \prod_{i=1}^{n_f} \{h_0(t_{fi}) e^{X_i\beta}\}^{\delta_{fi}} \times (-1)^{d_f} \phi^{(d_f)}\left(\sum_i H_0(t_{fi}) e^{X_i\beta}\right)$$

where $\varphi(s)$ is the Laplace transform of the frailty distribution $g(z)$, $\varphi^{(d)}(s)$ is the d th derivative of $\varphi(s)$ with respect to s, and $d_f = \sum_{i=1}^{n_f} \delta_{fi}$.

The Laplace transform of the frailty distribution and its dth derivative have the following form

$$\phi(s) = \int_0^{\infty} e^{-zs} g(z) dz,$$

$$\phi^{(d)}(s) = (-1)^d \int z^d e^{-zs} g(z) dz.$$

To be more specific, the Laplace transform of the gamma frailty distribution and the Weibull baseline functions can be written as:

$$\varphi(s) = (1 + s/k)^{-k}$$

$$\phi^{(d)}(s) = (-1)^d \frac{(\kappa + d - 1)!}{\kappa! \kappa^{d-1}} (1 + s/\kappa)^{-\kappa - d},$$

where $s = \Sigma_i (\lambda t_i)^{\rho} e^{X_i\beta}$

Ascertainment correction

Family data are often collected through an affected individual, who is called the proband, and a correction for sampling bias needs to be applied to get unbiased parameter estimates. The likelihood is corrected by the probability of being ascertained through the proband who is affected by her or his age at examination a_{fp} [20,21] which can be expressed as:

$$P(T_{fp} < a_{fp} \mid x_{fp}) = \int \{1 - S(a_{fp} \mid z_f)\} g(z_f) dz_f$$

$$= 1 - \int e^{-z_f H_0(a_{fp}) e^{X_{fp}\beta}} g(z_f) dz_f$$

$$= A_f(\theta), \quad (4)$$

where p indexes the proband.

The ascertainment-corrected likelihood arising from n families can be obtained by dividing each family's likelihood contribution as in (3) by its probability of being ascertained as in (4), given as:

$$L^c(\theta) = \prod_{f=1}^{n} \frac{L_f(\theta)}{A_f(\theta)},$$

and the corresponding ascertainment-corrected log-likelihood can be expressed as:

$$\ell^c(\theta) = \sum_{f=1}^{n} \{\ell_f(\theta) - \alpha_f(\theta)\},$$

where $l_f(\theta) = \log L_f(\theta)$ and $\alpha_f(\theta) = \log A_f(\theta)$. Thus, we obtain the maximum likelihood estimator of the parameters $\theta = (\beta, \rho, \lambda, k)$ involved

in the model, including the frailty parameter k by maximizing above ascertainment-corrected likelihood. Also, the age-specific penetrance estimate is obtained using the cumulative distribution function $F(t|X) = P(T < t|X)$ in equation (2).

Missing genotypes

It is common that family data include missing genetic information. A modified segregation-based method can account for missing genotype information [3,19,20,]. It allows inference on missing genotypes by observed genotypes within families based on Mendelian transmission probabilities and genealogical relationships. Then the likelihood of a family is calculated by

$$L_f(\theta) = \sum_G L_f(\theta \mid G)P(G), \quad (5)$$

where the summation is over all possible genotype combination G of family members with missing genotypes given observed genotypes within the family with respect to the corresponding genotype probability $P(G)$. Here, the genetic probability $P(G)$ is determined by using Hardy-Weinberg equilibrium and Mendelian transmission probability in dependence on observed genotypes of parents or siblings. This segregation-based method for handling missing genotypes was implemented in R [22].

Robust variance estimator

For consistent variance estimation, the variance matrix of the parameters θ is obtained by the sandwich estimator

$$Var(\hat{\theta}) = I_o(\theta)^{-1} J(\theta) I_o(\theta)^{-1},$$

where $J(\theta)$ is the expected information matrix and $I_0(\theta)$ is the observed information matrix, they can be obtained by

$$U(\theta) = \sum_f \frac{\partial \ell_f(\theta)}{\partial \theta} - \sum_f \frac{\partial \alpha_f(\theta)}{\partial \theta};$$

$$J(\theta) = U^{\mathrm{T}}(\theta)U(\theta); \quad I_0(\theta) = -\frac{\partial^2 \ell^c(\theta)}{\partial \theta^{\mathrm{T}} \partial \theta}.$$

Robust variance estimator for penetrance: An asymptotic variance estimator for age-specific penetrance $F(t;\hat{\theta})$ is obtained using the Delta method

$$\mathrm{Var}\{F(t;\hat{\theta})\} = D_\theta^{\mathrm{T}}(t) Var(\hat{\theta}) D_\theta(t),$$

where $D_\theta(t)$ is the vector of partial derivatives of $F(t;\theta)$ and $Var(\hat{\theta})$ is the robust variance estimator for parameters θ.

Simulation Study

Family data generation

The simulation of family follows the principles described in Choi et al. [20]. We consider unclear families of size four---two parents and their two offspring, one of whom is the proband (an affected individual from whom the family is selected). At the first stage, all family members' ages at examination using a normal distribution with mean age 65 for the first generation and 45 for the second generation, with variance fixed at 2.5 years for both generations. It resulted in an average of 20 years difference between the parents and offspring. At the next stage, the proband's genotype of a major gene was determined conditioning on the proband's affection status by her/his age at examination, assuming Hardy-Weinberg equilibrium (HWE) and fixed population allele frequencies. The proband is required to be a mutation carrier of the major gene. Given the proband's genotypes, the genotypes of the other family members were then determined using HWE and Mendelian transmission probabilities calculated with Bayes' formula. To incorporate familial correlation, a proband's frailty Z was generated conditional on ascertainment via the proband being affected before her/his age at examination, i.e.,

$$Z \sim Z^* \mid T < a_{\mathrm{p}},$$

where Z^* assumed to follow a Gamma distribution with mean 1 and variance 1/k and T to follow a Weibull model. We let the generated frailty value be shared among family members. The distribution of the frailty conditional on the proband being affected before the age at examination was derived in Appendix. Once we simulated the age at examination, genotype information for all family members and the shared frailty for the family, then the time-to-onset of individual i was simulated from the shared frailty model,

$$h(t_i|z,x_i) = h_0(t_i) z \exp(\theta x_i),$$

where x_i indicates if the ith individual of the family is a carrier of disease mutation gene, z represents the frailty value shared within the family and the baseline hazard was assumed to follow the Weibull distribution which has a form, $h_0(t) = \lambda\rho\{\lambda(t-20)\}^{\rho-1}$, assuming the minimum age of 20 years at onset.

The proband's age at onset was generated conditioning on the fact that the proband was affected before his(her) age at examination, a_p,

$$T_p \sim T \mid T < a_{\mathrm{p}}.$$

For the rest of family members, their times to onset were generated unconditionally. We also assumed the maximum age for followup was 90 years of age. Finally, the affection status, δ_i for individual i was determined by comparing the age at onset, T_i, and age at examination, a_i; δ_i=1 if $T_i < a_i$, and 0 otherwise.

Simulated scenarios

Data were simulated under different configurations. Each configuration, we simulated 500 random samples of 1000 families each, which are similar to the available sample sizes from many familial cancer registries.

We assumed Weibull baseline hazard functions with scale (λ) and shape (ρ) parameters equal to 0.012 and 3. This leads to a cumulative risk (i.e. 1 - survival probability) of 19% in the non-carrier group by age 70. Two penetrances were considered: high and low penetrances correspond to the log relative risk of a major gene (β) equals to 2 and 1, respectively. The high penetrance represents the lifetime risk of 80% by age 70 among carriers of a major gene, which assumes a rare gene (allele frequency=0.02) under the dominant model. The low penetrance provides the lifetime risk of 44% among common gene carriers (allele frequency=0.3) under the recessive model.

We considered two sources of familial correlation: one induced by 1) a frailty and the other by 2) a second gene variation. For case 1), the frailty parameter k took the values 2, 4 and 8 that could be regarded as high, medium and low intra-familial correlation. Simulated data were evaluated using the family-specific frailty model and the independent model which assumes no familial correlation as comparison. The bias and precision of log relative risk and penetrance estimates from the frailty and independent models were summarized in Tables 1 and 2.

For case 2), the second gene effect set to β_2=1.6 and 0.7 for large and small familial correlations, respectively, and the second gene was generated by assuming an unknown second gene under the dominant model with the allele frequency 0.2. The simulated data were fitted using the independent, frailty and two-gene models. The two-gene model assumes the presence of an unknown second gene as an addition to the major gene in the model. The bias and precision of log relative risk and penetrance estimates from the independent, frailty and two-gene models were summarized in Tables 3 and 4.

Simulation results

Shared frailty: Two modeling approaches---frailty and independent models were compared for different penetrance values (high and low) and for different magnitudes of familial dependence (*k=2,4,8)*. We summarized our simulation results in Tables 1 and 2 for the estimation of the log relative risk (β) and the penetrance by age 70 in terms of average bias ×100 (Bias%), robust standard error (SE) and root mean square error (RMSE).

Log relative risk estimation: Table 1 presents the summary of the simulation results in terms of accuracy and precision for estimating the log relative risk (β). The family-specific frailty model led to accurate estimates of the log relative risk for both high and low penetrances regardless of the *k* values; the magnitude of the bias ranged from −0.52% to 0.50% using the frailty model, whereas the independent model slightly underestimated the log relative risk; higher familial correlation (*k=2*) yielded more severe bias (Bias%= −15.1 and -5.22 under high and low penetrances, respectively). The magnitude of the bias was in general smaller than the SEs, which ranged from 0.085 to 0.099 in our settings. Although the frailty model yielded more reliable estimates of the log relative risks for both high and low penetrance settings, the independent model performed more efficiently in the presence of small familial correlation as expected; RMSEs from the independent model and frailty model were 0.125 vs. 0.131 for high penetrance and 0.131 vs 0.134 for low penetrance when *k=8*.

Penetrance estimation: The simulation results in terms of accuracy and precision for estimating the penetrance by age 70 among carriers were summarized in Table 2. Similar to the relative risk estimation, the family-wise frailty model also yielded accurate penetrance estimates (Bias%= −0.09 ~ −0.58). The independent model also led to substantial bias in penetrance estimates (Bias% = 16.42 ~ 4.21) for both high

		Frailty Model			Independent Model		
Penetrance	*k*	Bias%	SE	RMSE	Bias%	SE	RMSE
High	2	0.47	0.099	0.138	-15.10	0.085	0.181
	4	0.50	0.099	0.135	-8.49	0.087	0.143
	8	0.46	0.097	0.131	-4.43	0.088	0.125
Low	2	-0.19	0.095	0.130	-5.22	0.089	0.130
	4	-0.52	0.097	0.132	-3.13	0.094	0.129
	8	-0.12	0.098	0.134	-1.57	0.097	0.132

Bias% = Bias×100

Table 1: (Shared frailties) Estimating the log relative risk (RR) of the major gene effect: Bias, robust standard error (SE) and root mean square error (RMSE), comparison of frailty and independent modeling approaches.

		Frailty Model			Independent Model		
Penetrance	*k*	Bias%	SE	RMSE	Bias%	SE	RMSE
High	2	-0.57	0.067	0.090	16.42	0.021	0.166
	4	-0.58	0.057	0.079	9.06	0.022	0.094
	8	-0.43	0.050	0.068	4.84	0.023	0.056
Low	2	-0.09	0.059	0.079	14.60	0.028	0.149
	4	-0.10	0.054	0.073	7.85	0.028	0.084
	8	-0.10	0.050	0.066	4.21	0.027	0.054

Bias% = Bias ×100

Table 2: (Shared frailties) Estimating penetrance by age 70 among mutation carriers: Bias, robust standard error (SE) and root mean square error (RMSE) in comparison of frailty and independent modeling approaches.

High Penetrance (β_1 = 2)						
	β_2 = 0.7			β_2 = 1.6		
Model	Bias%	SE	RMSE	Bias%	SE	RMSE
Independent	-33.80	0.080	0.348	-54.75	0.082	0.553
Frailty	-4.99	0.097	0.136	-23.41	0.100	0.259
2-gene model	-7.90	0.115	0.170	-14.22	0.110	0.198
Low Penetrance (β_1 = 1)						
	β_2 = 0.7			β_2 = 1.6		
Model	Bias%	SE	RMSE	Bias%	SE	RMSE
Independent	-12.82	0.076	0.156	-24.36	0.070	0.254
Frailty	-2.36	0.085	0.116	-13.48	0.078	0.163
2-gene model	0.44	0.099	0.133	-2.11	0.092	0.126

Bias% = Bias×100

Table 3: (Second gene variation) Estimating the log hazards ratio of major gene effect in the presence of a second gene: Bias, robust standard error (SE) and root mean square error (RMSE) in comparison of the three models.

High Penetrance (β_1 = 2)						
	β_2 = 0.7			β_2 = 1.6		
Model	Bias%	SE	RMSE	Bias%	SE	RMSE
Independent	7.14	0.014	0.073	9.14	0.010	0.092
Frailty	-19.64	0.072	0.210	-8.70	0.045	0.101
2-gene model	1.90	0.073	0.086	3.64	0.036	0.057
Low Penetrance (β_1 = 1)						
	β_2 = 0.7			β_2 = 1.6		
Model	Bias%	SE	RMSE	Bias%	SE	RMSE
Independent	17.13	0.025	0.173	18.86	0.020	0.190
Frailty	-9.72	0.062	0.120	-1.37	0.047	0.065
2-gene model	13.21	0.109	0.186	12.12	0.073	0.147

Bias% = Bias×100

Table 4: (Second gene variation) Estimating the penetrance by age 70 among carriers in the presence of a second gene: Bias, robust standard error (SE) and root mean square error (RMSE) in comparison of the three models.

and low penetrances. The penetrance was more overestimated in the presence of higher familial correlation. In terms of RMSE, the frailty model provided more efficient penetrance estimates for family data with high to medium familial correlation than the independent model for both high and low penetrances. For example, RMSE=0.09 for the frailty model, 0.166 for the independent model when *k=2* (high familial correlation) under high penetrance.

Use of frailty model in the presence of a second gene variation: We suppose that there exists a second gene, other than a major gene shared within families to induce familial correlation, instead of frailty. This second gene was considered as completely unobserved. For modeling heterogeneity due to the unknown second gene, the shared frailty model was applied to take into account the familial correlation due to the second gene for estimating the log hazards ratio of the major gene (Table 3) and penetrance (Table 4). In addition, the simulated data with the two genes were fitted using the two-gene model which assumes the presence of an unknown second gene (dominant) as an addition to the major gene in the model.

The bias and efficiency of log hazards ratio and penetrance estimates from the three model (independent, frailty and two-gene models) were summarized in Table 3. The frailty model provided relatively accurate estimates of the major gene in the presence of the small second gene variation (β_2=0.7); they were slightly negatively biased about 5% and 2% for high and low penetrances, respectively. However, when a large second gene effect was present (β_2=1.6), both the frailty and independent models substantially underestimated the major gene effect (Bias= −13%~ −54%) whereas the two-gene model performed well in general providing relatively accurate estimates of the major gene under low penetrance.

For estimating the penetrance, Table 4 shows different models worked differently for high and low penetrances, depending on the second gene variations. The two-gene model provided the most accurate and efficient penetrance estimates under the high penetrance (β_1=2) among the three models while the frailty model performed well in terms of both accuracy and precision under the low penentrance (β_1=1). Interestingly, in the presence of larger second gene variation (β_2=1.6), the frailty model worked better providing less bias and smaller SEs.

Application to an Early Onset Breast Cancer Study

The data

We applied our approach to a family study of early-onset breast cancer among BRCA1/2 mutation carriers. The goal is to estimate the genetic relative risk and penetrance associated with mutations in the BRCA1 and BRCA2 genes separately.

The family data were collected from three population-based breast cancer family registries (Ontario, Northern California and Australia) as a part of the NCI-funded Breast Cancer Family Registries initiatives [23]. In this study we focus only on the early breast cancer families whose probands were affected before the age 40.

A total of 1505 early breast cancer families were identified by the three registries including 974 families genotyped for either BRCA1 or BRCA2. For BRCA1 analysis, we used 924 families (including 334, 248, and 342 families from Australia, Ontario, and Northern California, respectively) after exclusion of BRCA2 positive families in order to remove any possible confounding in the baseline risk estimation. Similarly, 876 families were used for BRCA2 analysis (321, 225, and 330 families from Australia, Ontario, and Northern California, respectively) after exclusion of BRCA1 positive families.

Results

The family data sampled from three population-based breast cancer registries were fitted into the proposed frailty model and independence model. The log relative risk and penetrance estimates for BRCA1 and BRCA2 are summarized in Table 5. The missing genotypes were inferred using a segregation-based approach described in equation (5), based on the observed genotype information available within families.

The parameters were estimated in two stages. First, the population allele frequency was estimated independently from the other parameters, using observed genotypes of founders. Then, considering it fixed, the other parameters were estimated by maximizing the ascertainment-corrected likelihood. The minor allele frequencies for BRCA1 and BRCA2 genes were 0.034 and 0.017, respectively. These estimates might be slightly overestimated because of ascertainment, but these estimates were close to published allele frequencies for BRCA1 and BRCA2 in USA [24], see (Table 2).

For BRCA1, the relative risk of breast cancer was estimated at $e^{1.11}$ = 3.03 (SE=1.08) under the independent model and $e^{1.212}$ = 3.36 (SE=0.64) incorporating the presence of familial correlation. For BRCA2, the relative risks were 3.47 (SE=1.13) with the independence assumption and 3.87 (SE=0.67) with the familial correlation.

Among BRCA1 mutation carriers, the lifetime risk estimates of breast cancer at age 70 were 0.61 (SE=0.12) ignoring the familial dependence and 0.64 (SE=0.07) with familial correlation. Also, among BRCA2 mutation carriers, they were 0.67 (SE=0.11) and 0.65 (SE=0.06).

Based on these estimates, BRCA2 gene appeared to have higher risks than BRCA1 gene both relatively and absolutely. Importantly, we notice that the relative risks were estimated slightly larger with higher precision (smaller SEs) when the familial correlation was taken into account compared to when it was ignored whereas the lifetime risks did not show this pattern; the shared frailty model provided higher lifetime risk estimates for BRCA1 carriers than the independent model but both models produced similar lifetime risk estimates for BRCA2 carriers.

In addition, we obtained the Akaike information criterion (AIC) values to compare these three modeling approaches shown in Table 6. The share frailty model yielded the smallest AIC values which demonstrates better fit than the independent model.

Summary and Conclusion

Familial time-to-event data arising from genetic studies include several features. First, family data are correlated. Second, the sampling of family data is often based on complex study designs. Third, data include some missing genetic information. In this paper, we investigated a family-specific frailty modeling approach for analyzing correlated time-to-event data to account for familial correlation and introduced an ascertainment-corrected likelihood approach to take the study design into account. Last, a modified segregation-based method enabled us to infer missing genotype data based on the observed genotype information provided within families.

In our statistical framework, we assumed a relatively simple ascertainment correction based on single ascertainment (which conditions on the observed age and genotype of the actual proband), where families were ascertained with probability proportional to the number of affecteds [19,21,25] and the observed pedigree structure was independent of who were the probands [26]. Since the probands have been identified through population-based cancer registries in the first stage of the study and their relatives selected only in a second stage, the single ascertainment seems reasonable. However, some families might also have been ascertained because they included multiple affected probands. In that situation, a more appropriate correction for ascertainment would be to condition on the probability that the family include at least one affected proband. However, calculating the exact ascertainment probability is computationally challenging as it involves a complex summation over all possible genotypes, phenotypes and frailties of the whole family. Alternatively, some forms of weighted likelihood have been proposed to handle complex ascertainment schemes within the Breast cancer family registries, where the weights are the inverse sampling probabilities for families and parameter estimates are obtained through a pseudo-likelihood [26] or a composite likelihood approach [2,27]. With frailty models, the computation of these likelihoods are also challenging but we are currently working on this extension.

	Log Relative Risk		Penetrance by age 70	
Model	BRCA1 (SE)	BRCA2 (SE)	BRCA1 (SE)	BRCA2 (SE)
Independent	1.109 (0.357)	1.243 (0.325)	0.609 (0.116)	0.665 (0.109)
Frailty	1.212 (0.189)	1.354 (0.173)	0.639 (0.067)	0.646 (0.063)

Table 5: Log relative risk and penetrance estimates of BRCA1 and BRCA2 genes.

Model	BRCA1	BRCA2
Independent	11584	10999
Frailty	11566	10978

Table 6: AIC values for comparison of different modeling approaches.

Our simulation study examined their consistencies and efficiencies for estimating relative and lifetime risks of a major gene, under both high penetrance with dominant model and low penetrance with recessive model. We found that the family-specific frailty model performed well for estimating both relative and lifetime risks in the presence of high to moderate familial correlation. However, when the familial correlation was weak, the independent model provided good results.

Our early onset Breast Cancer study also demonstrated the importance of incorporating the familial correlation in the analysis of correlated time-to-event data. The frailty-based likelihood approach was effectively implemented for modeling familial correlation for family data from population-based family registries. For future research, we can extend our modeling approach with univariate frailty distribution to use of multivariate frailty distributions for correlated frailties to better accommodate more complicated familial correlation structure.

Acknowledgements

This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. This work was also supported by the United States National Cancer Institute, National Institutes of Health under RFA # CA-95-003 as part of the Breast Cancer Family Registries (Breast CFR), and through cooperative agreements with Breast CFR and principal investigators from Cancer Care Ontario (UO1 CA69467), Northern California Cancer Center (U01 CA69417) and University of Melbourne (U01 CA69638). The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the CFRs, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government or the CFRs. The author is grateful to the reviewers for their constructive comments, especially to Dr. Thomas whose comments and suggestions were led to marked improvements in preparing this paper for publication.

References

1. Carayol J, Bonaïti-Pellié C (2004) Estimating penetrance from family data using a retrospective likelihood when ascertainment depends on genotype and age of onset. Genet Epidemiol 27: 109-117.
2. Choi Y, Briollais L (2011) An EM composite likelihood approach for multistage sampling of family data. Statistica Sinica 21: 231-253.
3. Kraft P, Thomas DC (2000) Bias and efficiency in family-based gene characterization studies: conditional, prospective, retrospective, and joint likelihoods. Am J Hum Genet 66: 1119-1131.
4. Vaupel JW, Manton KG, Stallard E (1979) The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography 16: 439-454.
5. Aalen OO (1992) Modelling heterogeneity in survival analysis by the compound Poisson distribution. Ann Appl Probab 4: 951-972.
6. Aalen OO, Tretli S (1999) Analyzing incidence of testis cancer by means of a frailty model. Cancer Causes Control 10: 285-292.
7. Hougaard P (1984) Life table methods for heterogeneous populations: distributions describing the heterogeneity. Biometrika 71: 75-83.
8. 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: 141-151.
9. Hougaard P (1986) Survival models for heterogeneous populations derived from stable distributions. Biometrika 73: 387-396.
10. Oakes D (1982) A model for association in bivariate survival data. JR Statist Soc B 44: 414-422.
11. Choi YH, Matthews DE (2005) Accelerated life regression modeling of dependent bivariate time-to-event data. Can J Stat 33: 449-464.
12. Hougaard P (2000) Analysis of multivariate survival data. Springer, New York.
13. Lawless JF (2003) Statistical models and methods for lifetime data. (2ndedn), Wiley- Interscience, New York.

14. Liang KY, Self SG, Chang YC (1993) Modelling marginal hazards if multivariate failure time data. JR Statist Soc B 55: 441-453.

15. Liang KY, Self SG, Bandeen-Roche KJ, Zeger SL (1995) Some recent developments for regression analysis of multivariate failure time data. Lifetime Data Anal 1: 403-415.

16. Pickles A, Crouchley R (1995) A comparison of frailty models for multivariate survival Data. Stat Med 14: 1447-1461.

17. Keiding N, Andersen PK, Klein JP (1997) The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates. Stat Med 16: 215-224.

18. Gong G, Whittemore AS (2003) Optimal designs for estimating penetrance of rare mutations of a disease-susceptibility gene. Genet Epidemiol 24: 173-180.

19. Thomas DC (2004) Statistical methods in genetic epidemiology. Oxford University Press, New York, USA.

20. Choi YH, Kopciuk KA, Briollais L (2008) Estimating disease risk associated with mutated genes in family-based designs. Hum Hered 66: 238-251.

21. Epstein MP, Lin X, Boehnke M (2002) Ascertainment-adjusted parameter estimates revisited. Am J Hum Genet 70: 886-895.

22. R development core team (2011) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

23. John EM, Hopper JL, Beck JC, Knight JA, Neuhausen SL, et al. (2004) The Breast Cancer Family Registry: an infrastructure for cooperative multinational, interdisciplinary and translational studies of the genetic epidemiology of breast cancer. Breast Cancer Res 6: R375-R389.

24. Fackenthal JD, Olopade OI (2007) Breast cancer risk associated with BRCA1 and BRCA2 in diverse populations. Nat Rev Cancer 7: 937-948.

25. Ewens WJ (1991) Ascertainment biases and their resolution in biological surveys. Handbook of Statistics 8: 29-61.

26. Vieland VJ, Hodge SE (1995) Inherent intractability of the ascertainment problem for pedigree data: a general likelihood framework. Am J Hum Genet 56: 33-43.

27. Chen L, Hsu L, Malone K (2009) A frailty-model based approach to estimating the age-dependent penetrance function of candidate genes using population-based case-control study designs: an application to data on the BRCA1 gene. Biometrics 65: 1105-1114.

28. Whittemore AS, Halpern J (1997) Multi-stage sampling designs in genetic epidemiology. Stat Med 16: 153-167.

This article was originally published in a special issue, **Advances in Markov Chain Monte Carlo Methods and Survival Analysis** handled by Editor(s). Dr. Faming Liang, Texas A&M University, USA; Dr. Nengjun Yi, University of Alabama at Birmingham, USA; Dr. Wenqing He, University of Western Ontario, Canada; Dr. Liuquan Sun, Institute of Applied Mathematics, Academy of Mathematics and Systems Science, China