Published in final edited form as:

J Am Stat Assoc. 2013 December 1; 108(504): 1205–1215.

Frailty Models for Familial Risk with Application to Breast Cancer

Malka Gorfine^{1,2},

Faculty of Industrial Engineering and Management, Technion - Israel Institute of Technology, Technion City, Haifa 32000, Israel

Li Hsu², and

Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109-1024, U.S.A.

Giovanni Parmigiani

Department of Biostatistics and Computational Biology, Dana Farber Cancer Institute Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, U.S.A.

Malka Gorfine: gorfim@ie.technion.ac.il; Li Hsu: lih@fhcrc.org; Giovanni Parmigiani: gp@jimmy.harvard.edu

Abstract

In evaluating familial risk for disease we have two main statistical tasks: assessing the probability of carrying an inherited genetic mutation conferring higher risk; and predicting the absolute risk of developing diseases over time, for those individuals whose mutation status is known. Despite substantial progress, much remains unknown about the role of genetic and environmental risk factors, about the sources of variation in risk among families that carry high-risk mutations, and about the sources of familial aggregation beyond major Mendelian effects. These sources of heterogeneity contribute substantial variation in risk across families. In this paper we present simple and efficient methods for accounting for this variation in familial risk assessment. Our methods are based on frailty models. We implemented them in the context of generalizing Mendelian models of cancer risk, and compared our approaches to others that do not consider heterogeneity across families. Our extensive simulation study demonstrates that when predicting the risk of developing a disease over time conditional on carrier status, accounting for heterogeneity results in a substantial improvement in the area under the curve of the receiver operating characteristic. On the other hand, the improvement for carriership probability estimation is more limited. We illustrate the utility of the proposed approach through the analysis of BRCA1 and BRCA2 mutation carriers in the Washington Ashkenazi Kin-Cohort Study of Breast Cancer.

Keywords

familial risk prediction; frailty model; multivariate survival; ROC analysis; risk index; breast cancer

1 Introduction

With progress in our understanding of the role of inherited genetic variation in determining susceptibility to a variety of illnesses, it has become increasingly critical to provide an accurate assessment of familial risk. In cancer, as well as elsewhere, personalized

¹To whom correspondence should be addressed.

²These authors contributed equally to this work

management of prevention strategies at the population level has shown the ability to increase survival in high-risk groups, while decreasing cost and complications in low-risk groups. For example, women carrying deleterious mutations of the BRCA1 and BRCA2 genes have 8-10 fold increase in the risk of developing breast cancer. They are counseled about various prophylaxis options, including prophylactic mastectomy, chemoprevention, and more intensive screening. However, these interventions are less often recommended to women who do not carry a known familial mutation, even though some may also have high risk for breast cancer.

Two aspects of prediction are generally considered in assessing genetic risk: assessing the probability of carrying an inherited high-risk genetic mutation; and predicting the absolute risk of developing diseases over time for those whose mutation status is known (Parmigiani et al. 1998b). These two components provide comprehensive information on risk for individuals who seek genetic counseling and will be the focus of this work.

Various methods for assessing carriership probabilities are available and they may be classified into two groups (Parmigiani et al., 2007): Mendelian or empirical. The Mendelian models are based on the conditional distribution of the phenotype given the genotype, the marginal distribution of the genotypes, and Mendel's law for the joint distribution of the family members' genotypes (Murphy and Matalik, 1969). In contrast, in the empirical approach, a training sample of genotyped individuals is used for direct modeling of the conditional distribution of the genotype given the phenotype (Couch et al., 1997; Frank et al., 1998; Hartge et al., 1999; Vahteristo et al., 2001). In breast cancer, comprehensive comparisons show that the Mendelian approach generally outperforms the empirical approach for carriership prediction in major cancer familial syndromes (Barcnas et al. 2006; Parmigiani et al. 2007; Antoniou et al. 2008). The Mendelian approach can also be extended to predicting disease risk for individuals who have not developed the disease yet (Parmigiani et al., 1998a). Since such approach fully uses the family history information, it likely outperforms the empirical approach in this situation as well. Therefore, we focus on Mendelian models. Two Mendelian-based models, BRCAPRO (Parmigiani et al. 1998) and BOADICEA (Antoniou et al. 2008) have been found to be the best performing Mendelian models for carrier-status prediction (see Barcnas et al., 2006; Parmigiani et al., 2007; Antoniou et al., 2008). To the best of our knowledge, no comparison has been published for the risk prediction of developing the disease over time.

An important facet of risk estimation is heterogeneity of risk among families, whether or not a high-risk mutation is present. This heterogeneity arises from genetic and environmental factors that are either unobserved or not yet sufficiently well understood. For example, in breast cancer, known mutations only account for a fraction of familial cases and there remains a significant amount of residual correlation even after accounting for those known genes (Claus et al., 1998; Antoniou et al., 2001). Graber-Naidich et al. (2011) also observed a substantial risk heterogeneity after adjusting for BRCA1/2 mutations in the National Institute of Child Health and Human Development's Women's Contraceptive and Reproductive Experiences (CARE) study. A similar result was found in The Women's Environment, Cancer, and Radiation Epidemiology (WeCARE) study (Begg et al., 2008) and population-based studies of breast and ovarian cancer (Antoniou et al. 2008). This translates into a substantial variation in the probability of developing the disease. Therefore, the accuracy of disease prediction may be substantially affected if the residual correlation is not properly accounted for.

The frailty model provides a natural approach to accounting for risk heterogeneity. A frailty in this context is a family-specific variate which is shared by all family members, acting multiplicatively on the hazard function. Frailties vary from family to family and follow a

distribution whose variance is governed by heterogeneity of disease risk, after accounting for known mutations and possibly other risk factors. Frailty distributions considered in the literature include gamma, positive stable, inverse Gaussian, compound Poisson, log-normal and others. See Hougaard (2000), Therneau and Grambsch (2000), and Duchateau and Janssen (2008) for reviews.

This article provides novel frailty-based genetic risk estimation procedures. There are several advantages to the frailty approach. First, a frailty can represent both unaccounted genes and environmental/lifestyle factors shared by the family members. Second, it is interpretable as a family-specific risk effect, which is appealing for individualized genetic risk estimation. In addition, the frailty model also allows for a relatively straightforward extension of existing open source software packages, particularly the BayesMendel R package (Chen et al., 2004) that includes BRCAPRO model (Berry et al., 1997 and Parmigiani et al., 1998) and others. This allows one to take advantage of the existing data structure and of extensive programs for complex pedigrees and multiple phenotypes. Through this widely distributed package, our work has the potential for an immediate concrete translation into genetic counseling and cancer prevention.

We propose two approaches for risk prediction: a marginalized approach based on integrating over the unknown frailty variate to evaluate the marginal survival function; and a conditional approach based on predicting the unknown frailty variate for each family. For the conditional approach, we propose a second stage to correct for underestimation of the familial risk. We conducted a systematic comparison of these approaches. The proposed methods for risk prediction and the results are also applicable in other situations where heterogeneity needs to be taken into account such as subject heterogeneity in recurrence event times and center heterogeneity in a multi-center clinical trials.

The rest of the paper is organized as follows. Section 2 presents the notation and the models. Section 3 presents our proposals for genetic risk estimation along with sampling-based pointwise confidence intervals. Section 4 summarizes the results from our extensive simulation study for contrasting the proposed approach with our previously established BRCAPRO model. In Section 5 the utility of the methods is illustrated by using the Washington Ashkenazi Kin-Cohort Study. Section 6 provides some concluding remarks.

2 Notation and model definition

For a given family, let T_j^o and C_j be the failure and censoring times of the j th member, for the health outcome of interest. Here $j = 0, 1, \dots, m$. Also, let G_j be 1 if member j is a carrier of a genetic susceptibility variant and 0 otherwise. Denote by π the frequency of high-risk variants, and assume a dominant effect so that at the population level $\Pr(G = 1) = \pi^2 + 2\pi(1 - \pi)$. This specification would need to be altered depending on the mode of inheritance of the gene. Let ω be a random variable representing the family-specific “frailty”, shared by all members of a family. The variation of frailties across families is described by known density function $f(\omega) = f(\omega; \theta)$, where θ is an unknown dependence parameter, which quantifies risk heterogeneity across families. Conditional on ω and the carrier status of the family members, $G = (G_0, G_1, \dots, G_m)$, the failure times of the family members are assumed independent. This extends the standard assumption made in mendelian model of conditional independence of phenotypes given genotypes. We will consider two models: the extended Cox (1972) proportional hazards model

$$\lambda^{(1)}(t|G_j, \omega) = \lambda_0(t) \exp(\beta G_j + \omega), \quad j=0, 1, \dots, m \quad (1)$$

and the stratified extended Cox model

$$\lambda^{(2)}(t|G_j, \omega = \lambda_{0G_j}(t) \exp(\omega), G_j = 0, 1, j = 0, 1, \dots, m) \quad (2)$$

where β is a regression coefficient, λ_0 is an unspecified conditional baseline hazard function, and $\lambda_{00}, \lambda_{01}$ are unspecified conditional baseline hazard functions among non-carriers and carriers of high-risk mutations, respectively. Model (1) is the popular multiplicative frailty model (see Hougaard, 2000; and reference therein) with a proportional hazards assumption for the mutation effect. However, the effect of BRCA1/2 mutations on breast cancer is known not to be proportional on the hazard function (Parmigiani et al., 1998b), which motivates model (2). Let also $T_j = \min(T_j^o, C_j)$, $\delta_j = I(T_j^o \leq C_j)$, $\Lambda_0(t) = \int_0^t \lambda_0(u) du$ and $\Lambda_{0g}(t) = \int_0^t \lambda_{0g}(u) du$, $g = 0, 1$.

We refer to the counselee as the family member for whom we calculate predictions of outcome over time and carriership (if unknown), given her/his current personal and familial history. The counselee's index is $j = 0$. The relatives' history consists of m relatives and is denoted by $T_R = (T_1, \dots, T_m)$ and $\delta_R = (\delta_1, \dots, \delta_m)$. The carrier status of the relatives is $G_R = (G_1, \dots, G_m)$. We assume that the effect of carrier status is subject specific, namely $\Pr(T_j, \delta_j | G_F, \omega) = \Pr(T_j, \delta_j | G_j, \omega)$, $j = 0, 1, \dots, m$. In many genetic counseling applications G_R is unobserved and thus we focus on this situation; however, our methods can be easily adapted to the case where the genotype of one or more relatives is available. Also, let $T_F = (T_0, T_1, \dots, T_m)$ and $\delta_F = (\delta_0, \delta_1, \dots, \delta_m)$.

For both models, we will consider two different risk prediction methods: a marginalized approach and a conditional approach. We will use the log-normal frailty model where $\omega \sim N(0, \sigma^2)$ with $\theta = \sigma^2$. The proposed approaches can be easily applied under any other frailty distribution. For clarity of presentation of the main idea, we assume that the true value of the population parameters $\psi^{(1)} = (\pi, \beta, \theta, \Lambda_0)$ or $\psi^{(2)} = (\pi, \theta, \Lambda_{00}, \Lambda_{01})$ are available. In practice, these can be estimated from family studies by well-established statistical methods (e.g., Gorfine et al., 2006; Zeng and Lin, 2007 for cohort family study, and Shih and Chatterjee, 2002; Hsu et al., 2004; Chatterjee et al., 2006; Chen et al., 2009; Gorfine et al., 2009; Graber-Naidich et al., 2011 for case-control family study).

3 Genetic risk estimation using a frailty model

We describe risk estimators for a counselee, given an observed family history and models (1) or (2). These methods use the unknown frailty variate to account for the unmeasured genetic or environmental risk factors, that together with the known genes explain the familial aggregation. While both our new methods estimate the counselee's risk conditional on his/her observed family history, they take different approaches. In the conditional modeling, the family-specific frailty is estimated based on the family history information, and then the risk is assessed given the estimated frailty. In contrast, the marginalized approach estimates the posterior distribution of the frailty given the family history information and averages the risk over the frailty with respect to the posterior distribution.

3.1 The marginalized approach

The marginalized approach uses the marginal survival function obtained by integrating the conditional survival function over the unknown frailty ω . Let $\mathcal{F}$ denote the observed information. For example, assume $\mathcal{F} = \{T_0, \delta_0 = 0, G_0, T_R, \delta_R\}$ so that G_0 is known but G_R is unknown. The risk of developing the disease at a future age t is $\Pr(T_0^o \leq t | \mathcal{F})$ for $t > T_0$.

Averaging over all possible configurations of the unknown genotypes of relatives we can write

$$\begin{aligned}\Pr(T_0^o > t | \mathcal{F}) &= \sum_{\zeta_R \in \{0,1\}^m} \Pr(G_R = \zeta_R | \mathcal{F}) \int \Pr(T_0^o > t | \mathcal{F}, \zeta_R, \omega) f(\omega | \mathcal{F}, \zeta_R) d\omega \\ &= \frac{\sum_{\zeta_R \in \{0,1\}^m} \Pr(G_R = \zeta_R | G_0) \int \Pr(T_0^o > t | G_0, T_R, \delta_R, \zeta_R, \omega) \Pr(T_R, \delta_R | \zeta_R, \omega) f(\omega) d\omega}{\sum_{\zeta_R \in \{0,1\}^m} \Pr(G_R = \zeta_R | G_0) \int \Pr(T_0^o > T_0 | G_0, T_R, \delta_R, \zeta_R, \omega) \Pr(T_R, \delta_R | \zeta_R, \omega) f(\omega) d\omega}\end{aligned}$$

Then, given model (M), $M = 1$ or 2 , and the conditional independence assumption given the frailty variate ω and the carrier status of family members (G_0, ζ_R),

$$\Pr(T_0^o > t | G_0, T_R, \delta_R, \zeta_R, \omega) \Pr(T_R, \delta_R | \zeta_R, \omega) = A_R^{(M)} e^{\omega \delta_R - \exp(\omega) \{h^{(M)}(t, G_0) + H^{(M)}(T_R, \zeta_R)\}}$$

where $A_R^{(1)} = e^{\beta \delta_R^T \zeta_R} \prod_{k=1}^m \lambda_0^{\delta_k}(T_k)$, $A_R^{(2)} = \prod_{k=1}^m \lambda_{0\zeta_{Rk}}^{\delta_k}(T_k)$, $\delta_R = \sum_{j=1}^m \delta_j$, $h^{(1)}(t, g) = \Lambda_0(t) \exp(\beta g)$, $h^{(2)}(t, g) = \Lambda_0(t)$, $H^{(M)}(T_R, \zeta_R) = \sum_{k=1}^m h^{(M)}(T_k, \zeta_{Rk})$, $M = 1, 2$, and ζ_{Rk} is the k th component of ζ_R .

Similarly, when the carrier status of the counselee is unknown and the goal is to estimate the probability of carriership, the observed information is $\mathcal{F}' = \{T_0, \delta_0 = 0, T_R, \delta_R\}$. The marginalized approach for estimation of the carrier probability of a counselee, given model (M) is to compute

$$\Pr(G_0 = 1 | \mathcal{F}') = \frac{\sum_{\zeta_R \in \{0,1\}^m} \Pr(G_F = (1, \zeta_R)) A_R^{(M)} \int e^{\omega \delta_R - \exp(\omega) \{h^{(M)}(t, 1) + H^{(M)}(T_R, \zeta_R)\}} f(\omega) d\omega}{\sum_{\zeta_F \in \{0,1\}^{m+1}} \Pr(G_F = \zeta_F) A_R^{(M)} \int e^{\omega \delta_F - \exp(\omega) \{H^{(M)}(T_F, \zeta_F)\}} f(\omega) d\omega}$$

Where $H^{(M)}(T_F, \zeta_F) = \sum_{k=0}^m h^{(M)}(T_k, \zeta_{Rk})$, $M = 1, 2$ and $\sum_{\zeta_F \in \{0,1\}^{m+1}}$ represents a sum over all possible configurations of the mutation carrier status of the m relatives and the counselee. Numerical procedures, such as Gaussian quadrature, can be used for solving the above integrals. The prior probabilities of carrier status of family members can be expressed as a function of the frequency π of high-risk variants, under Mendel's laws.

3.2 The conditional approach

This approach is based on replacing the unknown frailty of the counselee's family with its point estimate. For this we use the method of Ha et al. (2001) also identical to the penalized partial likelihood approach (Therneau and Grambsch, 2000). Specifically, we write the log-likelihood function based on the joint distribution $\Pr(T_F, \delta_F, \omega | G_F) = \Pr(T_F, \delta_F | G_F, \omega) f(\omega)$. Then, for model (M) and the log-normal frailty model, the estimation equation becomes

$$\sum_{j=0}^m \delta_j - \exp(\omega) \sum_{j=0}^m h^{(M)}(T_j, G_j) - \omega / \sigma^2 = 0. \quad (3)$$

When some genotypes are unknown, we replace each $\exp(\beta G_j)$ or $\Lambda_0 G_j$ for which $G_j, j = 0, \dots, m$ is unknown, by its expectation conditional on the observed information and ω . For example, if G_R is unknown and G_0 is known, we replace each unknown $\exp(\beta G_j), j = 1, \dots, m$ by $\exp(\beta) \Pr(G_j = 1 | \mathcal{F}, \omega) + \Pr(G_j = 0 | \mathcal{F}, \omega)$. Straightforward calculations yield that $\Pr(G_j = 1 | \mathcal{F}, \omega)$ equals

$$\frac{\sum_{\zeta_{R_j} \in \{0,1\}^{m-1}} \Pr(T_F, \delta_F | G_{R_j} = \zeta_{R_j}, G_j = 1, G_0, \omega) \Pr(G_{R_j} = \zeta_{R_j}, G_j = 1, G_0)}{\sum_{\zeta_R \in \{0,1\}^m} \Pr(T_F, \delta_F | G_R = \zeta_R, G_0, \omega) \Pr(G_R = \zeta_R, G_0)} \quad (4)$$

where R_j represents the remaining $m - 1$ relatives after excluding the j th member. Particularly, under model (1), $\Pr(G_j = 1 | \mathcal{F}, \omega)$ becomes

$$\frac{\sum_{\zeta_{R_j} \in \{0,1\}^{m-1}} e^{\beta(\delta_j + \sum_{k=1, k \neq j}^m \delta_k \zeta_{R_{jk}}) - \exp(w)\{h^{(1)}(T_j, 1) + H^{(1)}(T_{R_j}, \zeta_{R_j})\}} \Pr(G_{R_j} = \zeta_{R_j}, G_j = 1, G_0)}{\sum_{\zeta_R \in \{0,1\}^m} e^{\beta \delta_R^T \zeta_R - \exp(w) H^{(1)}(T_R, \zeta_R)} \Pr(G_R = \zeta_R, G_0)}$$

where $H^{(1)}(T_{R_j}, \zeta_{R_j}) = \sum_{k=1, k \neq j}^m h^{(1)}(T_k, \zeta_{R_{jk}})$ and $\zeta_{R_{jk}}$ is the k th element of ζ_{R_j} . The corresponding expression for model (2) can be derived similarly.

For solving (3), after the above mentioned replacements, one may use the Newton-Raphson algorithm. Specifically, if $\mathcal{F} = \{T_0, \delta_0 = 0, G_0, T_R, \delta_R\}$, $\hat{\omega}$ is defined as the root to the following equation:

$$\sum_{j=0}^m \delta_j - \exp(\omega) \left\{ \sum_{j=1}^m \tilde{h}^{(M)}(T_j, \cdot) + h^{(M)}(T_0, G_0) \right\} - \omega / \sigma^2 = 0 \quad (5)$$

where $M = 1$ or 2 , $\tilde{h}^{(1)}(T_j, \cdot) = \Lambda_0(t) \{ \exp(\beta) \Pr(G_j = 1 | \mathcal{F}, \omega) + \Pr(G_j = 0 | \mathcal{F}, \omega) \}$ and $\tilde{h}^{(2)}(T_j, \cdot) = \Lambda_{01}(t) \Pr(G_j = 1 | \mathcal{F}, \omega) + \Lambda_{00}(t) \Pr(G_j = 0 | \mathcal{F}, \omega)$.

A tempting approach would be to define a risk predictor at age $t > T_0$ based on $\Pr(T_0^o \leq t | \mathcal{F}, \omega)$ by replacing ω by $\hat{\omega}$. Continuing the above example where $\mathcal{F} = \{T_0, \delta_0 = 0, G_0, T_R, \delta_R\}$, we get

$$1 - \exp \left[-\exp(\hat{\omega}) \left\{ h^{(M)}(t, G_0) - h^{(M)}(T_0, G_0) \right\} \right], \quad (6)$$

for $M = 1, 2$. However, when the frequency of high-risk variants is low, as is the case with BRCA mutations, risk predictions based on (6) may be biased downwards, and underestimate the total number of events in the population. We will illustrate this later in the simulation study. To correct for this underestimation, we propose to calibrate the risk by considering ω as a risk index (Cai et al., 2010). Specifically, we estimate the probability $\Pr(T_0^o \leq t | \hat{\omega}, T_0^o \geq T_0, G_0)$ based on an estimator of $\mu_g(t, \hat{\omega}) = \Pr(T_0^o \leq t | \hat{\omega}, G_0 = g), g = 0, 1$ such as the non-parametric kernel Kaplan-Meier estimator (Beran, 1981; Dabrowska 1989, and others) or a Cox proportional hazards models with ω as a covariate. We recommend using an external training dataset, which can be the same one used for estimation all other

parameters. To summarize, our calibrated-conditional risk predictor at age $t > T_0$ consists of two stages: (i) estimate ω ; and (ii) plug-in $\hat{\omega}$ in the calibration model estimator to obtain $\{\hat{\mu}_{G_0}(t, \hat{\omega}) - \hat{\mu}_{G_0}(T_0, \hat{\omega})\} / \{1 - \hat{\mu}_{G_0}(T_0, \hat{\omega})\}$.

For the counselee carrier status prediction, for example with $\mathcal{F}' = \{T_0, \delta_0 = 0, T_R, \delta_R\}$, the estimator would be based on the calibration model and

$$\Pr(G_0=1|\mathcal{F}', \omega) = \frac{\sum_{\zeta_R \in \{0,1\}^m} \Pr(G_F=(1, \zeta_R)) \Pr(T_0^0 > T_0 | G_0=1, \omega) \sum_{j=1}^m \Pr(T_j, \delta_j | G_j=\zeta_{Rj}, \omega)}{\sum_{\zeta_F \in \{0,1\}^{m+1}} \Pr(G_F=\zeta_F) \prod_{j=0}^m \Pr(T_j, \delta_j | G_j=\zeta_{Rj}, \omega)}$$

such that $\Pr(T_j, \delta_j | G_j, \omega), j = 0, 1, \dots, m$ are estimated based on the calibrating model μ_{G_j} with ω .

Both the marginalized and the calibrated-conditional provide the counselor a practical tool for risk prediction, which uses external data (for estimation of $\psi^{(1)}$ or $\psi^{(2)}$ and the calibrating model) but also incorporates the dependency among family members. The genetic risk estimators can be updated easily when new data are added. The calibrated-conditional approach has two main advantages over the marginalized: (1) it does not require any numerical integration, and (2) it provided a quantitative summary $\exp(\omega)$ of the family's risk. On the other hand, the marginalized approach integrates over the distribution of ω , and thus accounts for additional uncertainty in the estimated frailty. Because of the conditional independence assumption given (ω, G_F) , both approaches can be incorporated in the BayesMendel R package (Chen et al., 2004) without major modifications.

3.3 Confidence Intervals on Predictions

The variance of the proposed estimators of both the survival probability and carrier status probability cannot be derived in closed form. We recommend a simple Monte Carlo procedure, similar to that of Parmigiani et al. (1998b), for estimating these variances. Assume that $\psi^{(\hat{M})}$, $M = 1, 2$, converges weakly to a Gaussian process, and that the variance function of $\psi^{(\hat{M})}$ can be estimated by $\sigma^{(M)}$; here the effective sample size is that of the dataset used for estimating $\psi^{(M)}$ and $\Sigma^{(M)}$. This is the case for each of the estimation methods listed at the end of Section 2. In the marginalized approach, and given model (M) , $M = 1$ or 2 : (i) Generate B draws of ψ 's from the multivariate normal distribution $N(\psi^{(\hat{M})}, \sigma^{(\hat{M})})$ (ii) For each calculate the counselee risk prediction (or carrier probability), denoted by $\psi^{(\hat{b})}$, $b=1, \dots, B$. (iii) Estimate the risk prediction (or carrier probability) variance by the empirical variance of $p^{(1)}, \dots, p^{(B)}$ and construct a Wald-type confidence interval, or simply evaluate empirical quantiles.

For the calibrated-conditional approach, the variability of the calibrating stage should be added as well. Assume that the calibrating stage is based on a consistent and asymptotically Gaussian process estimator, and denote its variance function estimator by $V(\hat{\omega})$. We propose to proceed as follows: (i) Generate B draws ψ 's from the multivariate normal distribution $N(\psi^{(\hat{M})}, \sigma^{(\hat{M})})$ (ii) For each $\psi^{(\hat{b})} b = 1, \dots, B$ calculate $\omega^{(\hat{b})}$ (iii) Given $\omega^{(\hat{b})}$, generate $\tilde{\mu}_{G_0}^{(b)}$ from the multivariate normal distribution $N(\mu_{G_0}(\omega), V(\omega;))$ and calculate the counselee risk prediction (or carrier probability) $P^{(\hat{b})}$, $b = 1, \dots, B$. (iv) Estimate the risk prediction (or carrier probability) variance by the variance of $P^{(\hat{1})}, \dots, P^{(\hat{B})}$ and construct a Wald-type confidence interval.

4 Simulation study

We next describe simulation results for comparing the breast cancer risk prediction of the current version of the BRCAPRO to those of our proposed extension, using settings that are faithfully representing hereditary breast cancer.

4.1 Simulation study design

Our simulation study consists of a random sample of 80,000 counselees with known carrier status, each providing information on the disease history, but not the genotypes, of their mother and two sisters. We use the log-normal frailty model with mean 0 and variance $\sigma^2 = 1, 1.5$ or 2, correspond to Kendall's τ correlations of 0.27, 0.34 and 0.39, respectively. Since our main interest is to compare the performances of the proposed approaches to that of BRCAPRO, the hazard function was determined based on the BayesMendel R package version 2.0-1 (Chen et al., 2006) which provides the marginal hazard functions (here marginal is with respect to the frailty variate), separately for carriers and non-carriers, at ages 1 through 110 years. Specifically, let $S_j^m(\cdot)$, $j=0, 1$, be the BayesMendel marginal survival functions of non-carriers and carriers, respectively. The conditional cumulative baseline hazard function at time t , $\Lambda_{0j}(t)$, $j=0, 1$, is the value of h that minimizes

$\{S_j^m(t) - \int_{-\infty}^{\infty} \exp(-he^x) \phi(x; \sigma^2) dx\}^2$ Where $\phi(\cdot; \sigma^2)$ is the normal density with mean 0 and variance σ^2 . This was done by numerical integration.

For each of the 80,000 counselees, the genotypes of the parents were generated with high-risk variant frequency $\pi = 0.01$, and the genotypes of their offsprings were generated based on Mendel's law of inheritance. Given the frailty variate ω , for each family member we then independently generated a random variable, Y , from the exponential distribution with parameter $\exp(\omega)$. Then, given the family member carrier status, G , $G = 0$ or 1, her/his failure time, T^0 , was back-calculated by equating $Y = \Lambda_{0G}(T^0)$. The censoring times were normally distributed with mean 55 and standard deviation 15. We excluded counselees with breast cancer diagnoses, and counselees older than 65 years of age, resulting in a sample of about 57,000 counselees (see in Table 1). For the calibrated-conditional approach, a separate large training dataset was generated to obtain the calibration model $\mu_g(\cdot, \cdot)$, $g = 0, 1$, by using a simple Cox proportional hazards models with ω as a covariate.

4.2 Results

Table 1 presents the results of predicting the risk of having breast cancer by ages $T_0 + 5$ and $T_0 + 7$, contrasting the predictions with the true event after a follow-up of 5 and 7 years, respectively. The proposed methods are based on the stratified extended Cox model (2). Numerical evaluation of integrals with no analytical closed form was done by using Gauss-Hermite quadrature with 20 nodes. The results of BRCAPRO are based on the assumption that given the counslee carrier status, the family history has no additional contribution in prediction. Hence, the probability of being free of breast cancer at age t , $t > T_0$, given $(G_0, T_0, \delta_0 = 0)$ becomes $S_j^m(t)/S_j^m(T_0)$ for carriers ($j = 1$) and non-carriers ($j = 0$). To provide an additional benchmark, we show the results which could be obtained if the true frailty variate ω was known. These are not practically achievable, as the frailty is never known in real applications.

The comparison between the methods is based on various measures: the area under the curve (AUC) of the receiver operating characteristic (ROC), the ratio of the observed and the expected number of events, the mean squared error of prediction, and the net reclassification improvement (NRI) (the NRI results are presented in the Supplementary Material). The AUC, or equivalently, the C index (Harrell et al., 1982) is an established performance index

for diagnostic tests and can be used for assessing the performance of various approaches in risk prediction at various ages. Specifically, ROC-AUC measures the probability that the predicted risk for a randomly selected diseased subject exceeds the risk prediction of a randomly selected non-diseased individual. Along with each ROC-AUC estimate, Table 1 presents the p-values of a two-sided hypothesis test for equality of the areas of each of our frailty-based methods compared to the BRCAPRO model. All proposed methods outperform BRCAPRO in predicting the risk, based on the ROC-AUC. The frailty-based methods (whether calibrated or not) are very similar in terms of ROC-AUC, except perhaps among carriers and at high values of σ^2 , where the conditional approach gives slightly better risk prediction than the marginalized approach. In general, all methods perform better in the whole set than in the carrier and non-carrier sub-groups, since each sub-group is more homogeneous and therefore more challenging to prediction algorithms. With regard to calibration (O/E), BRCAPRO, marginalized and calibrated-conditional perform equally well. All the p-values of a one-sided test that ROC-AUC > 0.5 were less than 0.05 (details not shown).

Table 2 presents the empirical mean squared error for prediction (MSEP) defined by

$n^{-1} \sum_{i=1}^n \{\hat{S}_i(t_i) - S_i(t_i)\}^2$ where n is the total number of counselees, S_i is the true survival value based on the true frailty variate, $S_i(t_i) = \exp[-\exp(\omega)\{\Lambda_{0G_i}(t_i) - \Lambda_{0G_i}(T_i)\}]$, and $\hat{S}_i$ is the estimated value. The results indicate that the frailty-based methods outperform BRCAPRO, except for the calibrated-conditional method in the case of low dependence among the family members ($\sigma^2 = 1$) presumably because of additional uncertainty introduced in the calibration step. However, as the dependency within family increases, the advantage of all the frailty-based methods in comparison with BRCAPRO increases, and the best approach is the marginalized approach.

The simulation results indicate that, in data where familial dependence beyond that given by the genotypes exists, our approaches can capture it effectively and produce predictions with improved discrimination. For additional insight, Table 3 reports the median estimated probability of each method stratifying by familial risk based on ω , and carrier status of counselee. Four categories were used for ω based on the quartiles of the distribution of ω . We used the data of Table 1 and present results of two settings, 5 years follow-up with $\sigma^2 = 1$ and 7 years follow-up with $\sigma^2 = 2$. Results for other settings are between those presented. The median estimated probability of BRCAPRO is nearly constant across quartiles of ω , both for carriers and non-carriers. In contrast, the median predicted probability of each of our two methods increases as ω increases. These results explain the differences in the ROC-AUC of BRCAPRO and our methods, and quantify how large the difference between these predicted methods could be. Table 5 shows, as an example, three counselees with their family history and carrier status, the value of ω , the event status at age $T_0 + 7$, and the estimated probability by each method.

In summary, the main conclusion from our simulation study is that we can expect tangible advantages in relaxing the conditional independence assumption made in BRCAPRO model by using the frailty approach.

Next we would like to provide an insight on the effect of missing relatives' carrier status given the carrier status of counselee. Table 5 contrasts the ROC-AUC and MSEP of 5- and 7-years of follow-up when carrier status of family members is known as well, to the case in which only the carrier status of counselee is observed (as in Tables 1-2). The ROC-AUC slightly increases in carriers, but not in non-carriers. This is probably due to the low carrier frequency: among true non-carriers, the probability of the mother and two sisters being all

non-carriers equals 0.98255, while the probabilities of each of the other seven possible scenarios is approximately 0.0025.

To illustrate the effect of family size, we also looked at the case in which each counselee provides information on her mother and eight sisters, and carrier status is known only for the counselee. As an example, if $\sigma^2 = 2$, out of 57,427 counselees 1016 developed the disease within 7 years. The ROC-AUCs of BRCAPRO, marginalized, calibrated-conditional, and with known frailty variate were: 0.672, 0.781, 0.762 and 0.868, respectively. The following are the corresponding numbers among non-carriers: 0.640, 0.764, 0.742, 0.860; and among carriers: 0.577, 0.702, 0.709, 0.767. Given the carrier status of a counselee, family size can have a strong effect on discrimination when predictions are made using the frailty model, as large pedigrees all of for more accurate estimation of family specific effects.

To demonstrate our method for constructing pointwise confidence interval, consider, for example, the marginalized approach, and model (1). The values of ψ (the superscript (1) is omitted) and the variance of ψ were determined based on a separate simulated dataset consisting of 2000 independent families, using the hazard rates of the BayesMendel package version 2.0-1. We used the EM-algorithm for estimation, in the spirit of Zeng and Lin (2007). 2500 counselees were sampled, and their survival probabilities $S(\psi)$ at 5 or 7 years after their current age were calculated based on ψ . For each counselee, a Wald-type confidence interval was constructed based on $B = 200$ bootstrap samples following the procedure in Section 3.3. The following results, based on 1829 counselees (after excluding diseased or over 65 years old subjects) with $\sigma^2 = 1.5$, summarize the empirical coverage rates based on the proportion of counselees in which their Wald-type confidence intervals covers $S(\psi)$. At nominal confidence levels of 90% and 95%, the empirical coverage rates were, respectively, 89.4% and 94.6% at 5 years, and 89.7% and 94.1% at 7 years. These results indicate that the empirical coverage rates are reasonably close to the nominal confidence levels. Similar results were observed for other values of σ^2 and hence are omitted.

Next we focus on the probability of carrying a high-risk variant. We consider the same simulation study design, with the modification that the carrier status of the counselee is unknown. We generated 57,274 counselees of which 817 are carriers. The ROC-AUC (p-value) for $\sigma^2 = 2$ are 0.677 (< 0.0001), 0.679 (< 0.0001), 0.641 (< 0.0001) for BRCAPRO, the marginalized and the conditional methods, respectively. The three methods are all similarly calibrated, with O/E ratio of 0.978, 0.991 and 1.004, respectively. The ROC-AUC figures are lower than typically reported in the literature because of the small family size considered here. Similar results were observed for other configuration hence results are not shown.

Overall, accounting for correlation among family members does not contribute significantly to the performance of predictions of carrier status, at least for this family size. This may be explained by the fact that risk heterogeneity among families, as formulated in models (1) and (2), models unobserved environmental factors and other genetic components that affect both carriers and non-carriers of the BRCA genes. Increasing the value of ω for a particular family will change the hazard of disease for both carriers and noncarriers by the same proportion. While the carrier probability calculation does depend on absolute risk, it is primarily driven by the density ratio of penetrances between carriers and noncarriers (Katki, 2006). Thus, variation in ω , as formulated here, has limited added value for predicting the carrier status of the known high-risk genes. In a way, whether or not accounting for risk heterogeneity has less impact on the weight. Therefore, BRCAPRO has similar performance to that of the methods that account for risk heterogeneity.

Our simulation study was performed using the true population-level parameters', values since the finite sample uncertainty in the parameters estimates is expected to affect all the methods similarly. However, the frailty-based methods may be affected additionally by the uncertainty on the frailty parameter estimate. We illustrate this effect by sampling from the log-normal frailty model with $\sigma^2 = 2$ while applying the frailty-based prediction method with $\sigma^2 = 2.4$ for the 7-years prediction. The following are the corresponding ROC-AUC results of the marginalized and the calibrated-conditional approaches: whole sample: 0.744, 0.736; non-carriers: 0.721, 0.713; and carriers: 0.619, 0.644. The corresponding O/E results are: whole sample: 0.953, 1.072; non-carriers: 0.938, 1.087; and carriers: 1.098, 0.967.

4.3 Frailty modeling and polygenic effects

In modeling familial aggregation that is not explained by known susceptibility genes, polygenic model are an important alternative to frailty models. For example, the carrier probability model BODICEA (Antoniou et al., 2008) uses a polygenic model to capture residual genetic effects. It is interesting to contrast the two approaches: while the polygenic model has points of contact with random effects models, it has a different dependency structure among the family members compared to our frailty models. Even for nuclear families, BOADICEA and our frailty approach are not equivalent. Specifically, our model assumes a random variable ω shared by all nuclear family members, with a one-dimensional variance parameter describing variation across families. The polygenic model used in BOADICEA assumes that each individual has a random effect and the joint distribution of the random effects is multivariate normal with an age-dependent covariance matrix that is governed by the kinship coefficients. Then, BOADICEA integrates over the polygenic random effects. This involves in a larger parameter dimension, which can be a limitation when parsimony is important. Also, the polygenic model captures the familial correlation according to genetic relationship, to capture inheritable effects. While this model can be more accurate than a frailty model if the unaccounted correlation is mostly genetic, it may be too restrictive here because the correlation of family members is not only due to shared genetics but also due to shared environment. For example, breast cancer mortality rates of Japanese women are far lower than U.S. rates when these women live in Japan, but closely approach those for whites when Japanese families are raised in the United States (Buell, 1973).

It would have been interesting to report comparisons of the two approaches in our simulated and real examples. Unfortunately logistic challenges arising from the way in which BOADICEA is presently available to the public made it impractical, as users do not have control on the input parameters necessary to implement a fair comparison. As this is likely to change, it will be important to address this comparison in future work.

5 The Washington Ashkenazi Study

To illustrate the utility of our proposed procedures, we consider the Washington Ashkenazi Kin-Cohort Study (WAS; Struwing et al., 1997). In this study, blood samples and questionnaires were collected from Ashkenazi Jewish men and women volunteers living in the Washington, DC area. The questionnaire included information on cancer and mortality history of the first-degree relatives of the volunteers. Volunteers were tested for specific mutations in BRCA1/2 genes, though their relatives were not. For this analysis, we consider female volunteers for whom we have data on breast cancer about their mothers and two sisters, totaling to 327 counselees and their families. Among the 327 counselees, 12 were BRCA1/2 mutations carriers, and 20 had breast cancer in the 20 years prior to year 1996 when the data were collected.

For each subject, we then constructed her breast cancer history at time T_0 , defined as the observed counselee age minus $c_0 = 20$ years. We used 20 years in order to have as large a number of breast cancer cases as possible. Then, we estimated the counselee carrier status probability at T_0 , and the probability of each counselee developing breast cancer during the following 20 years, starting at T_0 . We compared the observed data to the predictions obtained using various values of $\psi^{(2)}$ under the stratified Cox model (2). For the cumulative hazards Λ_{0g} , $g = 0, 1$, we used the hazards of the R package BayesMendel version 2.0-1. For the allele frequency of BRCA1/2 among Askenazi-Jewish we used two values: 0.01 based on Rao et. al. (1996) and 0.02 based on Struwing et al., (1997) and Hartge et al. (1999). Kendall's τ was assumed to be between 0.25 and 0.43 (Chen et al., 2009; Graber-Naidich et al., 2011) which under a log-normal frailty model implies σ^2 between 0.9 and 2.4.

Tables 6 and 7 provide the ROC-AUC, the ratio of the observed and the expected number of cases along with a 95% bootstrap confidence interval, and the Brier Score,

$n^{-1} \sum_{i=1}^n \{\delta_i - 1 + \hat{S}_i(t_i)\}^2$, for the risk of developing breast cancer within 20 years (Table 6), and for the carrier status (Table 7). For predicting risk over time, the calibrated-conditional frailty model performs better than BRCAPRO at all levels of σ^2 , while the marginalized approach is roughly equivalent. While sample sizes are too small to assess calibration accurately, all models appear to predict too few cases. Higher values of σ^2 tend to improve calibration but decrease discrimination. In this example, the calibrated-conditional method performs better than the marginalized approach in terms of ROC-AUC, although in the simulation study such situation was observed only among carriers and high values of σ^2 . In the calibrated-conditional method, the number of expected breast cancer events increases and the number of expected mutation carriers decreases with σ^2 . In general, the number of expected breast cancer events is expected to increase (decrease) with σ^2 among high (low) risk families, i.e. families with high (low) value of ω ; and the number of expected mutation carriers is expected to decrease with increasing σ^2 as this dependency parameter captures dependence among family members that is due to factors other than known mutations.

6 Discussion

Personalized risk assessment is playing an increasingly important role in medicine and is an active areas of development for statistical methods in biomedical research. A wide array of methods are being developed for incorporating the effect of genetic and environmental factors on disease risk. At the same time, our understanding of the nature of many common and complex diseases, including diabetes, cardiovascular disease, and most forms of cancer, is evolving. A pattern that is emerging from genome-wide association studies suggests the existence of a large number of small effects, and a potentially even large number of interactions. Similarly, understanding of the role of environmental exposures, dietary factor and other behaviors reveals a few important effects along with a plethora of complex and smaller ones, possibly mediated by the genetics. In this scenario, family history will remain a very powerful marker for a large number of yet unknown shared genetic and non-genetic effects that contribute to risk heterogeneity across families.

In this work we presented simple and efficient methods for including familial dependency in risk assessment. We showed in simulations and real data that these models are worthy of serious consideration. While our context is that of frailty model to account for risk heterogeneity, the general conclusion regarding the comparison of the marginalized approach and the (calibrated-) conditional approach can be expected to hold in other situation where heterogeneity needs to be taken into account. Examples include accounting for subject heterogeneity in predicting recurrence event times, or center heterogeneity in

predicting survival outcomes in multi-center studies. In this paper we focus on risk prediction by providing an estimate of the probability of developing disease by a future time t and the confidence interval for this estimate. In situations where uncertainty on classifying subjects into diseased and non-diseased is of interest, additional sampling variation needs to be incorporated.

Our work provides the foundation for several potentially useful extensions. We only considered first-degree relatives and made the simplifying assumption of identical dependency parameters for any pair of relatives. However, in practice, a counselee is usually providing a family history beyond the first-degree relatives and it is important to extend our methods to accommodate any family size and use a flexible dependency structure which reflects the precise relation between relatives. In such generalization, avoiding integration would be practically helpful, highlighting the potential of the calibrated-conditional approach. For example, consider the following generalization in which models (1) and (2) are replaced by $\lambda_0(t) \exp(\beta G_j + b^T Z_j)$ and $\lambda_{0G_j}(t) \exp(b^T Z_j)$, respectively, where b is a family-level random vector with multivariate density function $f(b; \Theta)$ indexed by parameters Θ , and Z_j is a set of covariates plus the unit component. The dimension of b can be defined by the desired dependency structure among the relatives to reflect consanguinity and unmeasured shared environmental exposures. A non-parametric maximum likelihood estimators for λ_0 , β , λ_{0G_j} and Θ can be found in Zeng and Lin (2007). Our proposed methods can be extended by replacing any univariate integration with respect to ω with the multivariate integration with respect to b , and by replacing the estimating equation (3) with estimating equations for b .

The models discussed here do not include covariates other than the mutation carrier status. However, when additional information on covariates is available, the frailty models can be extended to include them in a straightforward way. Another important extension is the consideration of competing risks and multiple disease outcomes, and can be found in Gorfine et al. (2013).

The main idea of the conditional approach is estimating the frailty variate based on (3) and then using $\Pr(T_0^o \leq t | \mathcal{F}, \hat{\omega})$ for risk prediction. An alternative approach for estimating ω could be to minimize the MSE. The best MSE predictor of $\exp(\omega)$ is the posterior mean of $\exp(\omega)$ (Bickel and Doksum, 2001, Theorem 1.4.1), namely

$$E\{\exp(\omega) | \mathcal{F}\} = \frac{\sum_{\zeta_R \in \{0,1\}^m} \Pr(G_R = \zeta_R | G_0) f \exp(\omega) \Pr(T_0^o > T_0 | G_0, T_R, \delta_R, \zeta_R, \omega) f(T_R, \delta_R | \zeta_R, \omega) f(\omega) d\omega}{\sum_{\zeta_R \in \{0,1\}^m} \Pr(G_R = \zeta_R | G_0) f \Pr(T_0^o > T_0 | G_0, T_R, \delta_R, \zeta_R, \omega) f(T_R, \delta_R | \zeta_R, \omega) f(\omega) d\omega}.$$

There is a resemblance between the marginalized approach and the posterior mean approach in the sense that the integration is with respect to $\exp(\omega)$. Additional simulations results not shown here in detail also indicate that both approaches give very similar results. For the conditional approach, it is easy to verify that if $\exp(\omega)$ follows the gamma frailty model with expectation 1 and variance σ^2 , the two approaches are identical. This is not the case under the normal frailty model. Moreover, the posterior mean approach is very well calibrated in simulation, and so the calibration stage may not be needed. For example, under the log-normal frailty model, $N(0, 2)$, and in the case of 5 years prediction, the ratios of the observed and expected number of events are: 1.094, 1.102 and 1.037 for the whole sample, non-carriers and carriers, respectively. However, the posterior mean approach requires integration, whereas the conditional approach does not.

In the Supplementary Material we present the effect of frailty model misspecification on predictions. We considered the case in which the true distribution of $\exp(\omega)$ is gamma but the log-normal frailty model is being used for prediction. In this specific setting we observed only a modest effect on the predicted values.

In summary, we hope to have provided a set of practical methodologies for frailty-based familial risk prediction. This allows to build on existing Mendelian models of susceptibility, and to generalize them in a way that can provide substantial practical improvement in risk assessment for families with clusters of disease. While the paper is focused on familial risk prediction, the methods developed here can be applied to other situations where frailty-based prediction is warranted.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Dr. Nilanjan Chatterjee for facilitating the access of the Washington Ashkenazi Jewish data. Work of Giovanni Parmigiani is supported by grant KG081303 from the Susan G Komen Breast Cancer Foundation and grants NIH/NCI 5P30 CA006516-46 and 1R21CA177233-01. Work of Li Hsu is supported by the following NIH grants P01CA53996, R01AG014358 and P50CA138293. Work of Malka Gorfine is supported by the Israel Science Foundation (ISF) grant 2012898.

The R code for running our methods can be provided upon request by the first author.

References

- Antoniou AC, Pharoah PDP, McMullan G, Day NE, Ponder BA, Easton D. Evidence for further breast cancer susceptibility genes in addition to BRCA1 and BRCA2 in a population-based study. *Genetic Epidemiology*. 2001; 21:1–18. [PubMed: 11443730]
- Antoniou AC, Cunningham AP, Peto J, et al. The BOADICEA model of genetic susceptibility to breast and ovarian cancers: updates and extensions. *British Journal of Cancer*. 2008; 98:1457–1466. [PubMed: 18349832]
- Barcenas CH, Hosain GMM, Arun B, Zong J, Zhou X, Chen J, Cortada JM, Mills GB, Tomlinson GE, Miller AR, Strong LC, Amos CI. Assessing BRCA carrier probabilities in extended families. *Journal of Clinical Oncology*. 2006; 24:354–360. [PubMed: 16421416]
- Begg CB, Haile RW, Borg A, et al. Variation of Breast Cancer Risk Among BRCA1/2 Carriers. *Journal of the American Medical Association*. 2008; 299:194–201. [PubMed: 18182601]
- Beran, R. Technical Report. Univ California; Berkeley: 1981. Nonparametric regression with randomly censored survival data.
- Berry DA, Parmigiani G, Sanchez J, Schildkraut J, Winer E. Probability of carrying a mutation of breastovarian cancer gene BRCA1 based on family history. *Journal of the National Cancer Institute*. 1997; 89:227–238. [PubMed: 9017003]
- Bickel, PJ.; Doksum, KA. *Mathematical Statistics, Basic Ideas and Selected Topics*. 2nd. Vol. Vil I. Prentice Hall; 2001.
- Buell P. Changing Incidence of Breast Cancer in Japanese–American Women. *J Natl Cancer Inst*. 1973; 51(5):1479–1483. [PubMed: 4762931]
- Cai T, Tian L, Uno H, Solomon S, Wei LJ. Calibrating parametric subject-specific risk estimation. *Biometrika*. 2010; 92:389–404. [PubMed: 23049123]
- Chen S, Iversen E, Friebe T, Finkelstein D, Weber B, Eisen A, Peterson L, et al. Characterization of BRCA1 and BRCA2 mutations in a large United States sample. *Journal of Clinical Oncology*. 2006; 24(6):863. [PubMed: 16484695]

- Chen S, Wang W, Broman K, Katki HA, Parmigiani G. BayesMendel: an R Environment for Mendelian Risk Prediction. *Statistical Applications in Genetics and Molecular Biology*. 2004; 3(1) Article 21.
- Chen L, Hsu L, Malone K. A frailty-model-based approach to estimating the age-dependent penetrance function of candidate genes using population-based case-control study design: an application to data on the BRCA1 gene. *Biometrics*. 2009; 65:1105–1114. [PubMed: 19210733]
- Chatterjee N, Kalaylioglu Z, Shih JH, Gail MH. Case-control and case-only designs with genotype and family history data: estimating relative risk, residual familial aggregation, and cumulative risk. *Biometrics*. 2006; 62:36–48. [PubMed: 16542227]
- Claus EB, Schildkraut J, Iversen ES, Berry D, Parmigiani G. Effect of BRCA1 and BRCA2 on the association between breast cancer risk and family history. *Journal of the National Cancer Institute*. 1998; 90(23):1824–1829. [PubMed: 9839523]
- Collaborative Group on Hormonal Factors in Breast Cancer. Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without the disease. *Lancet*. 2001; 358:1389–1399. [PubMed: 11705483]
- Couch FJ, DeShano ML, Blackwood MA, Calzone K, Stopfer J, Campeau L, Ganguly A, Rebbeck T, Weber BL. BRCA1 mutations in women attending clinics that evaluate the risk of breast cancer. *New England Journal of Medicine*. 1997; 336:1409–1415. [PubMed: 9145677]
- Cox DR. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society B*. 1972; 34:187–220.
- Dabrowska DM. Uniform consistency of the kernel conditional Kaplan–Meier estimate. *Annals of Statistics*. 1989; 17:1157–1167.
- Duchateau, L.; Janssen, P. *The frailty model*. Springer; New York: 2008.
- Frank T, Manley S, Olopade O, et al. Sequence analysis of BRCA1 and BRCA2: Correlation of mutations with family history and ovarian cancer risk. *Journal of Clinical Oncology*. 1998; 16:2417–2425. [PubMed: 9667259]
- Graber–Naidich A, Gorfine M, Hsu L, Malone K. Survival analysis of case-control family data with general semi-parametric shared frailty model and missing genetic information. *Lifetime Data Analysis*. 2011; 17:175–194. [PubMed: 21153764]
- Gorfine M, Zucker DM, Hsu L. Prospective survival analysis with a general semiparametric shared frailty model – a pseudo full likelihood approach. *Biometrika*. 2006; 93:735–741.
- Gorfine M, Zucker DM, Hsu L. Case-control survival analysis with a general semiparametric shared frailty model – a pseudo full likelihood approach. *Annals of Statistics*. 2009; 37:1489–1517. [PubMed: 19684881]
- Gorfine M, Hsu L, Zucker DM, Parmigiani G. Calibrated prediction for multi-variate competing risks models. To appear in *Lifetime Data Analysis*. 2013
- Hartge P, Struwing JP, Wacholder S, Brody LC, Tucker MA. The prevalence of common BRCA1 and BRCA2 mutations among Ashkenazi Jews. *American Journal of Human Genetics*. 1999; 64:963–970. [PubMed: 10090881]
- Ha ID, Lee Y, Song JK. Hierarchical likelihood approach for frailty models. *Biometrika*. 2001; 88:245–253.
- Harrell FE Jr, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *Journal of the American Medical Association*. 1982; 247:2543–2546. [PubMed: 7069920]
- Hougaard, P. *Analysis of Multivariate Survival Data*. New York: Springer; 2000.
- Hsu L, Chen L, Gorfine M, Malone K. Semiparametric estimation of marginal hazard function from case-control family studies. *Biometrics*. 2004; 60:936–944. [PubMed: 15606414]
- Katki HA. Effect of Misreported Family History on Mendelian Mutation Prediction Models. *Biometrics*. 2006; 62:478–487. [PubMed: 16918912]
- Leal SM, Ott J. A likelihood approach to calculating a risk support interval. *American Journal of Human Genetics*. 1994; 54:913–917. [PubMed: 8178830]
- McCulloch CE, Neuhaus JM. Prediction of random effects in linear and generalized linear models under model misspecification. *Biometrics*. 2011; 67:270–279. [PubMed: 20528860]

- Murphy EA, Mutalik GS. The Application of Bayesian Methods in Genetic Counseling. *Human Heredity*. 1969; 19:126–151.
- Parmigiani, G.; Berry, D.; Iversen, J.; Müller, P.; Schildkraut, J.; Winer, E. Gatsonis, C.; al, E., editors. Modeling Risk of Breast Cancer and Decisions about Genetic Testing; Case Studies In Bayesian Statistics. 1998a. p. 173–268.<http://ftp.isds.duke.edu/WorkingPapers/97–26.ps>
- Parmigiani G, Berry D, Aguilar O. Determining carrier probabilities for breast cancer–susceptibility genes BRCA1 and BRCA2. *American Journal of Human Genetics*. 1998b; 62:145–158. [PubMed: 9443863]
- Parmigiani G, Friebe T, Iversen ES, et al. Validity of models for predicting BRCA1 and BRCA2 mutations. *Annals of Internal Medicine*. 2007; 147:441–450. [PubMed: 17909205]
- Roa BB, Boyd AA, Volcik K, Richards S. Ashkenazi Jewish population frequencies for common mutations in BRCA1 and BRCA2. *Nature Genetics*. 1996; 14:185–187. [PubMed: 8841191]
- Shih JH, Chatterjee N. Analysis of survival data from case–control family studies. *Biometrics*. 2002; 58:502–509. [PubMed: 12229984]
- Struwing JP, Hartge P, Wacholder S, et al. The risk of cancer association with specific mutations of BRCA1 and BRCA2 among Ashkenazi Jews. *New England Journal of Medicine*. 1997; 336:1401–1408. [PubMed: 9145676]
- Therneau, TM.; Grambsch, PM. Modeling survival data: extending the Cox model. New york: Springer–Verlag; 2000.
- Vahteristo P, Eerola H, Tamminen A, Blomqvist C, Nevanlinna H. A probability model for predicting BRCA1 and BRCA2 mutations in breast and breast–ovarian cancer families. *British Journal of Cancer*. 2001; 84:704–708. [PubMed: 11237395]
- Zeng D, Lin DY. Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society B*. 2007; 69:507–564.

Table 1

BRCAPRO model versus the proposed frailty-based methods: ROC-AUC and the ratio of the observed and expected number of events under log-normal frailty model.

σ^2		n	n_1	BRCAPRO			Marginalized			Conditional			Calibrated-Conditional			Known frailty variate		
				ROC-AUC	O/E		ROC-AUC	B vs M	O/E	ROC-AUC	B vs C	O/E	ROC-AUC	B vs CC	O/E	ROC-AUC	O/E	
				<i>5 years</i>														
1.0	W	57186	765	0.667	1.039	0.701	0.0000	1.038	0.699	0.0000	1.450	0.697	0.0000	1.065	0.802	1.032		
	NC	56355	674	0.634	1.030	0.669	0.0000	1.028	0.667	0.0000	1.472	0.666	0.0000	1.092	0.785	1.024		
	C	831	91	0.575	1.113	0.644	0.0334	1.114	0.644	0.0330	1.309	0.640	0.0705	0.903	0.793	1.098		
1.5	W	57193	782	0.676	1.061	0.726	0.0000	1.060	0.726	0.0000	1.619	0.724	0.0000	1.070	0.840	1.054		
	NC	56359	690	0.641	1.054	0.699	0.0000	1.051	0.698	0.0000	1.661	0.697	0.0000	1.088	0.827	1.046		
	C	834	92	0.589	1.129	0.643	0.0861	1.130	0.643	0.0918	1.362	0.639	0.1590	0.949	0.775	1.117		
2.0	W	57194	772	0.677	1.048	0.748	0.0000	1.046	0.748	0.0000	1.697	0.745	0.0000	1.062	0.857	1.042		
	NC	56359	682	0.643	1.042	0.724	0.0000	1.037	0.724	0.0000	1.752	0.722	0.0000	1.078	0.846	1.033		
	C	835	90	0.580	1.095	0.637	0.0825	1.115	0.643	0.0598	1.370	0.652	0.0419	0.954	0.789	1.119		
<i>7 years</i>																		
1.0	W	57186	1093	0.668	1.021	0.702	0.0000	1.020	0.701	0.0000	1.417	0.697	0.0000	1.084	0.805	1.014		
	NC	56355	975	0.636	1.017	0.675	0.0000	1.016	0.673	0.0000	1.447	0.670	0.0000	1.112	0.789	1.012		
	C	831	118	0.589	1.047	0.641	0.0729	1.048	0.644	0.0612	1.211	0.638	0.1653	0.896	0.777	1.034		
1.5	W	57193	1082	0.670	1.010	0.721	0.0000	1.008	0.720	0.0000	1.529	0.715	0.0000	1.066	0.834	1.002		
	NC	56359	966	0.639	1.008	0.697	0.0000	1.005	0.696	0.0000	1.577	0.691	0.0000	1.087	0.822	1.000		
	C	834	116	0.593	1.025	0.622	0.3231	1.033	0.624	0.3114	1.219	0.628	0.3181	0.920	0.774	1.023		
2.0	W	57194	1101	0.675	1.027	0.744	0.0000	1.026	0.742	0.0000	1.648	0.737	0.0000	1.070	0.856	1.021		
	NC	56359	980	0.644	1.022	0.721	0.0000	1.018	0.720	0.0000	1.703	0.715	0.0000	1.085	0.846	1.013		
	C	835	121	0.582	1.069	0.624	0.1442	1.087	0.631	0.1036	1.302	0.645	0.0651	0.966	0.783	1.089		

n = total number of counselees; n_1 = number of observed events.

W - whole sample; NC - non-carriers; C - carriers; O/E - ratio of observed and expected number of events; B vs M - two-sided p-value for testing BRCAPRO vs Marginalized; B vs C - two-sided p-value for testing BRCAPRO vs Conditional; B vs CC - two-sided p-value for testing BRCAPRO vs Calibrated-Conditional.

Table 2
BRCAPRO model versus the proposed frailty-based methods: Mean Squared Error for
Prediction ($\times 1000$) under log-normal frailty model

σ^2	BRCAPRO	Marginalized	Conditional	Calibrated-Conditional
<i>5 years</i>				
1.0	0.344	0.290	0.312	0.373
1.5	0.529	0.418	0.457	0.495
2.0	0.704	0.536	0.593	0.629
<i>7 years</i>				
1.0	0.665	0.562	0.604	0.700
1.5	1.018	0.808	0.884	0.940
2.0	1.354	1.030	1.141	1.213

Table 3
BRCAPro model versus the proposed frailty-based methods: Median estimated probability under log-normal frailty model

familial risk group	carrier	BRCAPro	M	CC	Known
<i>5 years and $\sigma^2 = 1$</i>					
1	no	0.01210	0.01045	0.00712	0.00211
2	no	0.01210	0.01057	0.00721	0.00552
3	no	0.01210	0.01077	0.00734	0.01030
4	no	0.01210	0.01202	0.00823	0.02382
1	yes	0.11176	0.08332	0.06808	0.03095
2	yes	0.10859	0.08699	0.07062	0.07453
3	yes	0.10939	0.09187	0.07500	0.13880
4	yes	0.10374	0.12667	0.11324	0.23981
<i>7 years and $\sigma^2 = 2$</i>					
1	no	0.01798	0.01311	0.00684	0.00124
2	no	0.01798	0.01332	0.00692	0.00470
3	no	0.01798	0.01376	0.00719	0.01158
4	no	0.01699	0.01661	0.00886	0.03637
1	yes	0.15082	0.10228	0.07863	0.02688
2	yes	0.14716	0.10716	0.08034	0.09391
3	yes	0.15229	0.12246	0.09337	0.20956
4	yes	0.14118	0.20359	0.17910	0.39324

familial risk group - based on ω , from lowest to highest; carrier - carrier status of counselee; M- Marginalized approach; CC - Calibrated-conditional approach; Known - prediction based on the true ω .

Table 4
BRCAPRO model versus the proposed frailty-based methods: The estimated probabilities of three counselees

Data	ω	BRCAPRO	M	CC	event by $T_0 + 7$
$T_0 = 48$ $\delta_0 = 0$ $G_0 = 0$ $T_R = (87, 54, 53)$ $\delta_R = (0, 0, 0)$	-0.234	0.018	0.012	0.010	no
$T_0 = 35$ $\delta_0 = 0$ $G_0 = 1$ $T_R = (63, 30, 38)$ $\delta_R = (0, 0, 0)$	0.391	0.112	0.075	0.027	no
$T_0 = 60$ $\delta_0 = 0$ $G_0 = 0$ $T_R = (57, 47, 28)$ $\delta_R = (0, 1, 1)$	4.125	0.027	0.137	0.141	yes

T_R = observed age of mother and two sisters; δ_R = event indicators of mother and two sisters. M- Marginalized approach; CC - Calibrated-conditional approach; Known - prediction based on the true ω .

Table 5
The effect of missing relatives' carrier status given the carrier status of counselee

σ^2		Known		Unknown	
		Marginalized	Calibrated-Conditional	Marginalized	Corrected-Conditional
			5 years - ROC-AUC		
1.0	whole sample	0.701	0.697	0.701	0.697
	non-carriers	0.669	0.666	0.669	0.666
	carriers	0.655	0.638	0.644	0.640
1.5	whole sample	0.726	0.724	0.726	0.724
	non-carriers	0.699	0.697	0.699	0.697
	carriers	0.644	0.642	0.643	0.639
2.0	whole sample	0.748	0.745	0.748	0.745
	non-carriers	0.724	0.722	0.724	0.722
	carriers	0.646	0.651	0.637	0.652
5 years - MSEP ($\times 1000$)					
1.0	whole sample	0.284	0.311	0.290	0.373
1.5	whole sample	0.407	0.434	0.418	0.495
2.0	whole sample	0.518	0.555	0.536	0.629
7 years - ROC-AUC					
1.0	whole sample	0.703	0.697	0.702	0.697
	non-carrier	0.675	0.671	0.675	0.670
	carrier	0.662	0.643	0.641	0.638
1.5	whole sample	0.722	0.716	0.721	0.717
	non-carrier	0.698	0.692	0.697	0.691
	carrier	0.638	0.638	0.622	0.628
2.0	whole sample	0.744	0.736	0.744	0.737
	non-carrier	0.722	0.713	0.721	0.715
	carrier	0.641	0.651	0.624	0.645
7 years - MSEP ($\times 1000$)					
1.0	whole sample	0.550	0.613	0.562	0.700
1.5	whole sample	0.787	0.852	0.808	0.940

		Known		Unknown	
σ^2		Marginalized	Calibrated-Conditional	Marginalized	Corrected-Conditional
		0.996	1.086	1.030	1.213
2.0	whole sample				

WAS data analysis. The marginalized and calibrated-conditional frailty-based methods for breast cancer prediction: ROC-AUC (p-value), and the ratio of observed and expected number of events (95% bootstrap confidence interval). BRCAPRO results: ROC-AUC=0.621(0.035); O/E=1.51(0.83,2.20); Brier Score=0.0947.

Table 6

		Marginalized			Calibrated-Conditional		
π	σ^2	ROC-AUC(p-value)	O/E(95% CI)	Brier Score	ROC-AUC(p-value)	O/E(95% CI)	Brier Score
0.01	1.00	0.628(0.027)	1.44(0.78,2.09)	0.0956	0.649(0.013)	1.76(0.91,2.62)	0.0902
	1.50	0.625(0.031)	1.41(0.76,2.06)	0.0960	0.648(0.014)	1.64(0.84,2.43)	0.0920
	2.00	0.621(0.035)	1.38(0.74,2.02)	0.0964	0.644(0.015)	1.51(0.78,2.23)	0.0941
	2.25	0.621(0.034)	1.37(0.73,2.01)	0.0966	0.644(0.015)	1.42(0.73,2.11)	0.0960
	2.50	0.620(0.036)	1.36(0.72,2.00)	0.0967	0.643(0.016)	1.37(0.71,2.02)	0.0971
0.02	1.00	0.628(0.027)	1.45(0.79,2.11)	0.0956	0.648(0.013)	1.78(0.91,2.65)	0.0915
	1.50	0.625(0.031)	1.42(0.76,2.08)	0.0956	0.648(0.013)	1.66(0.86,2.47)	0.0915
	2.00	0.621(0.035)	1.40(0.75,2.05)	0.0957	0.646(0.014)	1.54(0.79,2.28)	0.0936
	2.25	0.621(0.035)	1.39(0.74,2.04)	0.0959	0.645(0.015)	1.49(0.77,2.21)	0.0943
	2.50	0.620(0.036)	1.38(0.73,2.03)	0.0959	0.644(0.015)	1.38(0.71,2.06)	0.0969

Table 7

WAS data analysis. The marginalized and calibrated-conditional frailty-based methods for mutation carrier prediction: ROC-AUC (p-value), and the ratio of observed and expected number of events (95% bootstrap confidence interval). BRCAPro results under $\pi = 0.01$; ROC-AUC=0.718(0.005); O/E=0.66(0.27,1.04); Brier Score=0.0736; under $\pi = 0.02$; ROC-AUC=0.719(0.005); O/E=0.48(0.22,0.75); Brier Score=0.0902

π	σ^2	Marginalized			Calibrated-Conditional		
		ROC-AUC(p-value)	O/E(95% CI)	Brier Score	ROC-AUC(p-value)	O/E(95% CI)	Brier Score
0.01	1.00	0.721(0.005)	1.41(0.66,2.17)	0.0567	0.728(0.004)	0.86(0.35,1.34)	0.0632
	1.50	0.720(0.005)	1.43(0.67,2.19)	0.0567	0.674(0.020)	0.97(0.41,1.50)	0.0599
	2.00	0.721(0.005)	1.44(0.65,2.23)	0.0567	0.663(0.028)	1.15(0.48,1.77)	0.0565
	2.25	0.721(0.005)	1.44(0.66,2.21)	0.0567	0.663(0.028)	1.27(0.52,1.95)	0.0553
	2.50	0.721(0.005)	1.44(0.65,2.22)	0.0567	0.665(0.027)	1.35(0.54,2.10)	0.0548
0.02	1.00	0.721(0.005)	0.81(0.38,1.24)	0.0726	0.728(0.004)	0.59(0.26,0.91)	0.0799
	1.50	0.721(0.005)	0.81(0.39,1.24)	0.0726	0.663(0.028)	0.64(0.29,0.98)	0.0759
	2.00	0.722(0.005)	0.81(0.37,1.25)	0.0728	0.663(0.028)	0.72(0.33,1.10)	0.0717
	2.25	0.721(0.005)	0.81(0.38,1.25)	0.0729	0.663(0.028)	0.78(0.34,1.18)	0.0699
	2.50	0.721(0.005)	0.81(0.37,1.25)	0.0730	0.665(0.026)	0.81(0.35,1.24)	0.0689