

Estimating Disease Risk Associated with Mutated Genes in Family-Based Designs

Yun-Hee Choi^a Karen A. Kopciuk^b Laurent Briollais^a

^aSamuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Ont., ^bDivision of Population Health and Information, Alberta Cancer Board, Calgary, Alta., Canada

Key Words

Penetrance function · Efficient study designs · Ascertainment correction · Likelihood methods · Age-at-onset

Abstract

Objective: Many clinical decisions require accurate estimates of disease risk associated with inherited gene mutations. While several family-based designs have been proposed, their relative advantages remain unclear. **Methods:** We considered four commonly-used family-based designs and evaluated their performance in terms of accuracy and efficiency under several genetic models via simulation studies. We also derived and assessed several ascertainment-corrected likelihood methods for analyzing the simulated data and real data from 12 HNPCC pedigrees from Newfoundland. **Results:** We found that the design efficiency depends on the question of interest: the clinic-based family design with random probands yields the most efficient estimate of genetic relative risks, whereas the population-based family design with mutation carrier probands provides the most efficient penetrance estimates. For a particular question, an ascertainment correction seems possible using regular likelihood methods but the presence of genetic heterogeneity due to a strong second gene effect can lead to some bias in the risk estimation. **Conclusions:** This work gives a general methodological framework for analyzing family-based designs in gene characterization studies and provides more ra-

tionale for the choice of an efficient design and an appropriate likelihood method to estimate the risk associated with an inherited gene mutation.

Copyright © 2008 S. Karger AG, Basel

Introduction

Advances in the identification and treatment of genetically transmitted diseases have led to an increased need for reliable estimates of genetic susceptibility risk. These estimates are used in clinical settings to identify individuals at increased risk of being a disease allele carrier as well as to define the age-specific probability of developing a particular disease given one is a carrier (penetrance). Because Mendelian disease genes are rare, family-based studies are usually employed to estimate disease risk since multiple case families likely harbour the disease gene and generally phenotype information is available.

Several family-based study designs can be used to estimate the disease risk associated with a gene mutation when onset varies with age. Gong and Whittemore [1] discuss two basic types of family-based sampling schemes – population-based and clinic-based designs. For population-based designs, families are ascertained for study inclusion based on affected family members (probands) who are randomly sampled from the disease population. The proband is usually genotyped to determine if he/she carries the disease risk gene and additional genotype and

phenotype data can then be collected from other family members. The proband's family members are assumed to be representative of the general study population. Both prospective and retrospective cohorts of identified mutation carriers can be created and penetrance can be estimated by fitting a parametric or semi-parametric function of age at onset to the full pedigree data using appropriate likelihood formulations [2, 3]. The ascertainment correction for population-based designs requires modeling the probability that the probands are affected before their age at examination in the likelihood formulation.

For clinic-based designs, on the other hand, families are ascertained for study inclusion based on having multiple affected family members in addition to the affected probands. Pedigrees with many cases are highly informative because they are more likely to carry the disease gene mutation, but typically have not been ascertained in any population-based manner. Studies based on these high-risk families can be effective for characterizing the prevalence and penetrance of the mutated gene, but appropriate ascertainment correction for the multiple affected family members is needed to obtain valid penetrance estimation for the general population.

Ascertainment of high-risk individuals as a function of their risk, referred to as 'size-bias' [4], introduces a form of bias that results in more extreme members of the population being sampled. This problem is more likely to occur in the presence of disease risk heterogeneity due to other genes or shared environmental factors. Kraft and Thomas [3] found that the genetic risk estimates for case-control data on sibships could be seriously biased if risk heterogeneity between families was present and methods assuming homogeneity in the risks were employed. Similarly, Gong and Whittemore [1] found that the presence of additional risk factors, such as a second gene, could result in an upward bias in risk estimates.

To date, there is no consensus as to which family-based design (high-risk or population) provides more accurate and efficient estimates of disease risk. Also unclear is which ascertainment correction should be applied. Thus, both the ascertainment correction and study design impact the bias and efficiency of gene characterization studies. The objectives of this paper are to: (1) develop appropriate ascertainment-corrected likelihood methods for these different family-based designs and (2) evaluate their relative performance in terms of bias and efficiency under several genetic models. Performance is evaluated using simulated and real data from a set of HNPCC (Hereditary Non-Polyposis Colorectal Cancer syndrome) pedigrees from Newfoundland.

Methods

In our study, we will focus on two variants each of the population-based and clinic-based family designs. We will consider families of three generations which include two parents, their two children (one of whom is the proband), and their grandchildren. The following four sampling criteria are used to ascertain families for study inclusion:

- POP = {Proband is affected}
- POP+ = {Proband is affected mutation-carrier}
- CLI = {Proband is affected and at least one parent and one sib are affected}
- CLI+ = {Proband is affected mutation-carrier and at least one parent and one sibling are affected}

Population-based designs correspond to ascertainment criteria POP and POP+ since their study entry requirements are based only on the affected proband who is randomly sampled from the disease population. Ascertainment criteria CLI and CLI+ correspond to clinic-based designs which have multiple disease occurrences among family members. Important to note is that ascertainment criteria for the POP+ and CLI+ designs include families who have at least one member (proband) who carries the mutated gene of interest.

For diseases caused by mutated genes, the phenotype or outcome of interest often varies in age at onset so time to an event such death or disease diagnosis is relevant. Thus, the penetrance function for the disease susceptibility gene is defined either by a survival function or by a cumulative risk function (one minus the survival function). If we denote the age at onset by T , the affection status at age of examination by δ , where $\delta = 1$ if the disease occurred before age at examination a , i.e., $T < a$, or 0, otherwise, then the phenotype is given by $D = (T, \delta)$. Penetrance is then defined by the cumulative risk of a disease up to age t associated with genotype G . It can be estimated by maximizing a likelihood function for family data, conditional on the way the families were ascertained.

Ascertainment-Corrected Likelihood Methods

Family data can be considered as arising from a retrospective cohort study and can be analyzed using various likelihood methods [5, Chap. 11]. Ascertainment-corrected likelihood approaches for family data have been developed by several authors, based on either prospective likelihood [6], retrospective likelihood where family selection depends on phenotypes only [3] or phenotypes and genotypes [7], partial likelihood for multistage designs [8, 9], and population-calibrated likelihood estimation [10]. However, the relative advantage of these methods in various designs and for time to onset data is not well known. In this paper, we propose three ascertainment-corrected likelihoods for survival data and compare them using the four family-based study designs: POP, POP+, CLI, and CLI+.

An ascertainment-corrected likelihood, L , arising from a sample of n independent families for each family size n_f has the general form

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

where L_f is the conditional likelihood of family f obtained by dividing its contribution N_f to the likelihood by the probability A_f of its being ascertained. For family f with n_f family members we define $D = (D_1, \dots, D_{n_f})$ and $G = (G_1, \dots, G_{n_f})$ as the vector forms that represent their phenotypes and genotypes, respectively.

The three ascertainment-corrected likelihoods we consider for the analysis of the age-at-onset data from family-based designs are the prospective, retrospective, and joint likelihoods. All three likelihoods condition on the ascertainment process and are similar in nature to the ones presented in Kraft and Thomas [3] for binary phenotype data. Specifically, the prospective likelihood is based on modeling the time-to-event data given the family members' genotypes, the retrospective likelihood is based on modeling the probability of family members' genotypes given their phenotypes, and the joint likelihood is based on the joint probability of their genotypes and phenotypes.

Prospective Likelihood

The ascertainment-corrected prospective likelihood contribution for family f of size n_f is

$$L_f = P(D|G,A) = \frac{P(A|D,G)P(D|G)}{P(A|G)} \propto \frac{P(D|G)}{P(A|G)},$$

where we assume that $P(A|D,G)$ is equal to 1 if the vector D qualifies for ascertainment, and 0 otherwise, and so is independent of the parameters of interest.

If we further assume that individuals' phenotypes are conditionally independent given their genotypes, then the numerator can be expressed as

$$P(D|G) = \prod_{i=1}^{n_f} P(D_i|G_i) = \prod_{i=1}^{n_f} h(t_i|G_i)^{\delta_i} S(t_i|G_i). \quad (2)$$

Here $h(t_i|G_i)$ and $S(t_i|G_i)$ represent the hazard and survivor functions for individual i in family f at time t_i given their genotype G_i , respectively. In the prospective likelihood method, the ascertainment correction is based solely on the probability of individuals being affected before their age at examination. Thus, the denominator $P(A|G)$ for the population-based designs (POP, POP+) depends on the probability that the proband is affected before her (his) current age at examination and can be written as

$$P(A|G) = P(T < a_p | G_p), \quad (3)$$

where a_p and G_p represent the proband's age at examination and genotype, respectively. For the clinic-based designs (CLI, CLI+), the ascertainment correction is determined by up to four individuals – two affected siblings (including the proband) and at least one affected parent. By the conditional independence assumption of the disease status given genotype, the denominator for the clinic-based designs is given by

$$P(A|G) = P(T < a_p | G_p)P(T < a_s | G_s)\{1 - P(T \geq a_f | G_f)P(T \geq a_m | G_m)\},$$

where indices p, s, f, m represent the proband, proband's sibling, father and mother, respectively. The 'at least' condition for one parent to be affected is incorporated by the complement of neither being affected, i.e., $1 - P$ (both parents are not affected by their ages at examination).

Retrospective Likelihood

In the extreme case, the retrospective likelihood is obtained by conditioning on all phenotypes in the family [3]. The correction can also be achieved, with less loss of efficiency, by conditioning only on those individuals involved in the ascertainment set [11, 12]. This method is called 'assumption-free ascertainment' because the ascertainment event is not explicitly modeled, but note that the ascertainment set still needs to be specified. Let Ω be the set of all possible genotypic configurations of those in the ascertainment set A and G_ω be the vector of genotypes for genotypic configuration ω . Then the likelihood contribution for family f is given by

$$L_f = P(G|D,A) = \frac{P(A|D,G)P(D|G)P(G)}{P(D,A)} \propto \frac{P(D|G)P(G)}{\sum_{\omega \in \Omega} P(D,A|G_\omega)P(G_\omega)},$$

where $P(D|G)$ in the numerator can be expressed using the same form as in equation (2), and

$$P(G) = \prod_{i=1}^{n_f} \begin{cases} P(G_i), & \text{if individual } i \text{ is a founder,} \\ P(G_i|G_{m_i}, G_{f_i}), & \text{if individual } i \text{ is a nonfounder} \end{cases}$$

$P(G_i)$ is based on Hardy-Weinberg Equilibrium (HWE) and depends on the population allele frequency, which is estimated. $P(G_i|G_{m_i}, G_{f_i})$ is obtained using the Mendelian transmission probability for individuals whose parents are in the pedigree. The denominator, $P(D,A)$, represents the probability of observing the phenotypes of individuals who qualify for ascertainment and $P(D,A|G_\omega)$ is the conditional probability given the vector of genotypes in configuration ω . Denote A as the set of family members involved in the ascertainment criteria. For family member $i \in A$, we have

$$P(D_i,A|G_i) = \begin{cases} P(T < a_i | G_i), & \text{if individual } i \text{ affected by time } a_i, \\ P(T \geq a_i | G_i), & \text{otherwise.} \end{cases}$$

Thus,

$$P(D,A) = \sum_{\omega \in \Omega} \prod_{i \in A} P(D_i,A|G_{\omega_i})P(G_{\omega_i}),$$

where ω_i is the genotypic configuration for subject i and the sum is over all possible genotypic configurations of those in the ascertainment set A .

Ascertainment for the population-based designs (POP, POP+) is based only on the proband, hence the denominator can be written as:

$$P(D,A) = \sum_{\omega \in \Omega_p} P(T < a_p | G_\omega)^{\delta_p} P(T \geq a_p | G_\omega)^{1-\delta_p} P(G_\omega),$$

where Ω_p is the set of all possible genotypic configurations for the proband. Ascertainment for design POP+ is based on the carrier proband and has the same form as in equation (3) in the prospective likelihood. Thus, the prospective and retrospective likelihoods provide the same maximum likelihood estimates of the model parameters for design POP+.

For the clinic-based study designs (CLI, CLI+), the denominator for family f which includes up to four affected members can be obtained by

$$\begin{aligned}
P(D, A) = & \\
& \sum_{\omega \in \Omega} P(T < a_f | G_{\omega_f})^{\delta_f} P(T \geq a_f | G_{\omega_f})^{1-\delta_f} P(G_{\omega_f}) \times \\
& P(T < a_m | G_{\omega_m})^{\delta_m} P(T \geq a_m | G_{\omega_m})^{1-\delta_m} P(G_{\omega_m}) \times \\
& P(T < a_p | G_{\omega_p}) P(G_{\omega_p} | G_{\omega_m}, G_{\omega_f}) P(T < a_s | G_{\omega_s}) P(G_{\omega_s} | G_{\omega_m}, G_{\omega_f})
\end{aligned}$$

where $G_\omega = (G_{\omega_p}, G_{\omega_m}, G_{\omega_f}, G_{\omega_s})$ includes all possible genotypes of the four individuals in the ascertainment set, and δ_f and δ_m indicate the affection status of the father and the mother, respectively. For CLI+, the sum in the denominator is taken over all possible genotypes, provided that the proband carries a mutated allele of the major gene.

Joint Likelihood

The ascertainment correction for the joint likelihood uses the weakest ascertainment condition, $P(A)$, compared to $P(A | G)$ for the prospective likelihood or $P(D, A)$ for the retrospective likelihood. Thus, as for binary disease outcomes, it should be the most efficient of the three likelihoods considered here [3]. The contribution from family f to the likelihood is of the form

$$\begin{aligned}
L_f = P(G, D | A) = & \\
\frac{P(A | D, G) P(D | G) P(G)}{P(A)} \propto & \frac{P(D | G) P(G)}{\sum_{\omega \in \Omega} \sum_{v \in \Upsilon} P(A, D_v | G_\omega) P(G_\omega)},
\end{aligned}$$

where Υ is the set of all possible phenotypic configurations for those family members in the ascertainment set and D_v is the vector of phenotypes included in phenotypic configuration v .

For the population-based designs (POP, POP+), ascertainment correction is via the probability of the affected proband only, i.e.,

$$P(A) = \sum_{\omega \in \Omega_p} P(T < a_p | G_\omega) P(G_\omega),$$

where a_p is the proband's age at examination and the sum is over all possible genotypes for the proband. For POP+, the sum is over the possible genotypes for the carrier proband. The denominator, $P(A)$, in the joint likelihood for the population-based designs (POP, POP+) is proportional to the same expression in the retrospective likelihood, resulting in the same maximum likelihood estimates (MLEs) of the model parameters for these two designs. In the POP+ design, all three likelihoods are proportional to each other and so provide the same MLEs in this design setting.

For the clinic-based designs (CLI, CLI+), the summation in the denominator is now over both the phenotypes and genotypes of individuals in the ascertainment set, \mathcal{A} , which can be expressed as

$$\begin{aligned}
P(A) = & \sum_{\omega \in \Omega} \sum_{v \in \Upsilon} P(A, D_v | G_\omega) P(G_\omega) \\
= & \sum_{\omega \in \Omega} \left\{ 1 - P(T \geq a_f | G_{\omega_f}) P(T \geq a_m | G_{\omega_m}) \right\} P(G_{\omega_f}) P(G_{\omega_m}) \times \\
& P(T < a_p | G_{\omega_p}) P(G_{\omega_p} | G_{\omega_m}, G_{\omega_f}) P(T < a_s | G_{\omega_s}) P(G_{\omega_s} | G_{\omega_m}, G_{\omega_f}),
\end{aligned}$$

where $D_v = (\delta_{v_f}, \delta_{v_m}, \delta_{v_p}, \delta_{v_s})$ can take one of the following three vector forms based on the family structure considered in our study: (1, 1, 1, 1), (1, 0, 1, 1), (0, 1, 1, 1). Thus, the sum over the possible phenotypic configurations, $v \in \Upsilon$, results in the probability that at least one parent and two sibs are affected by their ages at examination. The outer summation is over all possible genotype combinations, $\omega \in \Omega$, of the four family members. For study design CLI+, the sum in the denominator is over all possible genotype vectors, provided that the proband carries a mutated allele of the major gene.

It is worth noting that the prospective and joint likelihoods for the clinic-based study designs are directly modeling the ascertainment process to correct for the selection, whereas the retrospective likelihood implicitly corrects for ascertainment by conditioning on the observed phenotypes for those included in the ascertainment set (since selection only depends on the phenotypes) [7]. Furthermore, the prospective likelihood does not depend on knowing the allele frequency, q , whereas the retrospective and joint likelihoods need to estimate it along with the model parameters.

Simulation Study

The simulation study aims were to (1) assess bias and efficiency in risk estimation (relative and absolute risks) for three ascertainment-corrected likelihood methods across four family-based study designs, (2) investigate potential bias in risk estimation in the presence of genetic heterogeneity due to a second gene effect, and (3) evaluate the first two aims under different genetic models. The genetic models we considered are dominant and recessive models with either a rare gene with high penetrance or a common gene with low penetrance.

The simulated three-generation family structure involved ages at onset and examination, gender, genotypes of the major and second genes, and phenotypes for each family member according to the assumed age-specific risk model. The following proportional hazards (PH) model which adopted a Weibull distribution for the baseline function was used to generate the age at onset, t ,

$$h(t | G) = \lambda \rho \{\lambda(t - 20)\}^{\rho-1} \exp(\beta_s x_s + \beta_1 x_1 + \beta_2 x_2). \quad (4)$$

Here x_1 and x_2 are indicator functions for the mutated allele carrier status for the major gene and second gene, respectively, and x_s distinguishes between males ($x_s = 1$) and females ($x_s = 2$). This PH model for simulating time-to-onset data within families permits the modeling of residual familial correlation by incorporating additional risk factors that aggregate within families, such as genetic variation due to a second gene. Since the Weibull distribution includes constant, increasing or decreasing hazard functions, this distribution choice enables flexible modeling of the baseline hazard function.

The penetrance at age t was estimated assuming this age-dependent penetrance function

$$F(t; \theta) = 1 - S(t; \theta) = 1 - \exp[-\{\lambda(t - 20)\}^\rho e^{\beta_s x_s + \beta_1 x_1}],$$

using the maximum likelihood estimates of the unknown regression and baseline hazard parameters. The cumulative risk function, $F(t; \theta)$, is obtained from the complement of the survivor function, $S(t; \theta)$, which is adjusted for gender and carrier status. The penetrance estimates at age 70 were calculated for four different groups – male carriers, male non-carriers, female carriers, and

female non-carriers. Simulation study results were obtained by combining male and female carriers, which were used to estimate the overall lifetime penetrance. The corresponding model-based standard errors of penetrance estimates were approximated using the Delta method. Details are given in the Appendices 1 and 2.

Simulation of Family Data

In simulating data for the two population-based family study designs, POP and POP+, we adopted the same family structure consisting of three generations of family members – two parents and their two offspring, one of whom is the proband. Each offspring has a spouse and their children range in number from two to five. At the first stage, we simulated the family members' gender with equal probabilities of being male and female and their ages at examination using a normal distributions with mean age 65 for the first generation and 45 for the second generation, with variance fixed at 2.5 years for both generations. The result was an average of 20 years (variance 1 year) difference between the second and third generations. At the next stage, conditioning on her/his gender, age at examination and affection status, we determined the proband's genotypes for the two genes assuming HWE with fixed population allele frequencies.

We considered two models of inheritance – dominant and recessive. The mutant allele frequencies of the major and second gene were selected to reflect the observed values in our HNPPC data set and were based on previous studies [1, 13] (see Application Section). For study designs POP+ and CLI+, the proband must be a mutation carrier of the major gene. Given the proband's genotypes, the genotypes of the other family members are then determined using HWE and Mendelian transmission probabilities calculated using Bayes' formula. Once we have simulated the gender, genotype, and age at examination information for all family members, we simulated the time-to-onset of the phenotype. For the proband, using model (4) but conditioning on the fact that the proband is affected before his (her) age at examination, a_p , his/her time to onset is $T_p \sim T | T < a_p$. The rest of the family members' times to onset are generated unconditionally. We assumed the minimum age of onset is 20 years and the maximum age for follow-up is 90 years. Finally, the affection status, δ , is determined by comparing the ages at onset, T , and examination, a ; $\delta = 1$ if $T < a$, and 0 otherwise.

To generate family data under the clinic-based designs, CLI or CLI+, the same procedures were repeated systemically to obtain sufficient families that fulfill the ascertainment criteria. Generally, simulating clinic-based families takes about 10–70 times longer than simulating data for the population-based family designs, depending on the magnitude of the second gene effect.

Parameter Selection

Data were simulated under four genetic inheritance models: Model (1) dominant with rare high-penetrance disease allele (cumulative risk at 70: 82% for male carriers, 43% for female carriers, $q = 2\%$), Model (2) dominant with rare low-penetrance disease allele (cumulative risk at 70: 52% for male carriers, 21% for female carriers, $q = 2\%$), Model (3) recessive with common high-penetrance disease allele (cumulative risk at 70: 82% for male carriers, 43% for female carriers, $q = 30\%$), Model (4) recessive with common low-penetrance disease allele (cumulative risk at 70: 52% for male carriers, 21% for female carriers, $q = 30\%$). In all of these models, the non-carrier penetrances were set at 15% and 5% for

males and females, respectively. We considered two different values of the major gene effect, $\beta_1 = 2.35$, and 1.5, corresponding to high and low penetrance, respectively. The other model parameters were fixed at $\beta_s = -1.13$, $\lambda = 0.016$; $\rho = 3$. Three different values for the second gene effect, β_2 , were chosen to be 0, 0.7, or 1.6 in order to create no, small, or large second gene variation, respectively, and the variant allele frequency was set at 0.2. The penetrance functions for all genetic models with no second gene effect (i.e., $\beta_2 = 0$) are presented in figure 1.

For each scenario studied, we simulated 500 random samples of 100 families each, which is similar to the available sample sizes from many familial cancer registries. The simulated data were evaluated using the three different likelihood methods described above and assumed no second gene effect (homogeneity of risk). Results of the simulation study are based on empirical summary measures of bias, robust standard error (SE), asymptotic relative efficiency (ARE), and root mean square error (RMSE). Bias is defined here as the average difference between the estimated and simulated risk values. In the presence of a second gene effect, true penetrance was calculated as the expected penetrance over the second gene distribution based on the true model. The robust SE is obtained as the average model-based standard error estimates using the sandwich variance estimator (see Appendix 2).

The ARE of one likelihood, L_1 , to another, L_0 , is given by the ratio of the inverse asymptotic estimates for the variance of $\sqrt{n}(\hat{\theta} - \theta)$,

$$ARE = \frac{1/var_1}{1/var_0} = \frac{var_0}{var_1}.$$

Here var_0 and var_1 represent the asymptotic variances of the parameter θ for the likelihoods L_0 and L_1 , respectively. A value of $ARE < 1$ indicates that the likelihood L_0 is more efficient than L_1 for estimating θ and a value of the $ARE > 1$ indicates the converse situation. The tradeoff between bias and precision is measured using the RMSE, which is defined as

$$RMSE = \sqrt{bias^2 + var}.$$

Simulation Study Results

In this simulation study, we obtained the maximum likelihood estimates (MLEs) from three different likelihood methods based on four different genetic models. We then compared the study designs using the MLEs of the log relative risk and lifetime risk (penetrance at age 70) to assess their accuracy, precision and accuracy-precision tradeoff. In addition, we explored how the degree of familial correlation due to the presence of a second gene affects the bias and efficiency tradeoff in a dominant major gene model with a rare but highly penetrant allele across the different study designs and likelihood methods considered.

Simulations without Second Gene Variation

Log Relative Risks

Absolute values of bias were always less than $|0.08|$ and generally positive under the dominant inheritance model (left side of table 1). The study designs that had a genotyped proband always had less bias than their counterparts with no genotyped proband. Under recessive inheritance models, bias in the log RRs was less than $|0.06|$ except in the two clinical-based designs analyzed with either the retrospective or joint likelihood method, where the bias

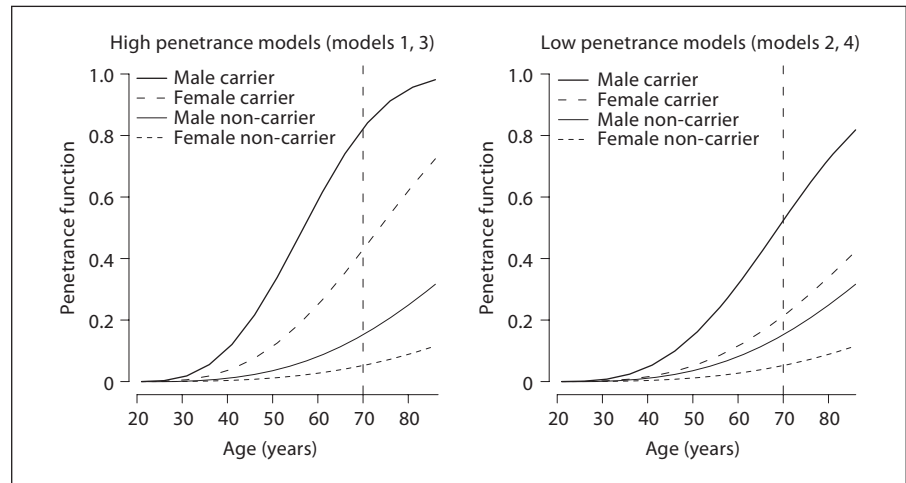


Fig. 1. Cumulative risk among four groups of male/female carriers/non-carriers, as specified by the high, low penetrance models when there is no second gene variation.

Table 1. Comparison of four genetic models in log relative risk (RR) estimates of the major gene effect: Bias, robust standard error (SE) and root mean square error (RMSE) from various study designs, using different likelihood methods

Method	Design	Dominant model with rare gene ($q = 2\%$)						Recessive model with common gene ($q = 30\%$)					
		high penetrance RR = 10.5			low penetrance RR = 4.5			high penetrance RR = 10.5			low penetrance RR = 4.5		
		Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE
Prospective	POP	0.03	0.35	0.35	-0.04	0.51	0.51	0.05	0.32	0.33	0.03	0.38	0.38
	POP+	0.04	0.34	0.34	0.03	0.36	0.36	0.05	0.34	0.34	0.05	0.36	0.36
	CLI	0.04	0.32	0.33	0.04	0.35	0.35	0.05	0.39	0.39	0.04	0.36	0.36
	CLI+	0.07	0.32	0.33	0.05	0.29	0.29	0.09	0.40	0.41	0.06	0.36	0.36
Retrospective	POP	0.07	0.25	0.26	0.09	0.31	0.33	-0.07	0.18	0.19	-0.01	0.20	0.20
	POP+	0.04	0.34	0.34	0.03	0.36	0.36	0.05	0.34	0.34	0.05	0.36	0.36
	CLI	0.05	0.17	0.17	0.04	0.16	0.17	0.14	0.18	0.23	0.09	0.14	0.16
	CLI+	0.01	0.24	0.24	0.01	0.21	0.21	0.35	0.32	0.48	0.35	0.22	0.41
Joint	POP	0.07	0.25	0.26	0.09	0.31	0.33	-0.07	0.18	0.19	-0.01	0.20	0.20
	POP+	0.04	0.34	0.34	0.03	0.36	0.36	0.05	0.34	0.34	0.05	0.36	0.36
	CLI	0.05	0.14	0.15	0.04	0.14	0.14	0.17	0.17	0.24	0.11	0.12	0.16
	CLI+	0.04	0.24	0.24	0.02	0.19	0.19	0.38	0.26	0.46	0.37	0.20	0.42

was substantial (between 0.11 and 0.37). The magnitude of the bias was in general much smaller than the standard errors (SEs), which ranged from 0.12 to 0.51 in all our settings. The level of penetrance had very little effect on biases and SEs. The precision of clinic-based designs was higher (smaller SEs) than the population-based designs in almost all our settings and among the clinic-based designs, the one with random affected probands (CLI) was the most efficient when fitted with retrospective and joint likelihoods. We also compared the efficiencies of the different likelihood methods. In most of the settings we considered, the retrospective likelihood method was almost as efficient as the joint likelihood (their AREs were close to or equal to 1) and the prospective likelihood was substantially less efficient. Under design POP+, all three likelihoods were proportional to each other

since the ARE was equal to one and so yielded the same maximum likelihood and corresponding asymptotic variance estimates. The root mean square error (RMSE) was mainly determined by the magnitude of the SE in all our settings. For all genetic models and likelihood methods, the clinic-based designs had RMSEs one third to one half of those obtained from population-based designs. The only exception was when the designs were analyzed with the prospective likelihood but this estimation method was the least efficient.

Penetrance Estimates

Absolute values of bias tended to be smaller for the penetrance estimates than for the log relative risks (i.e. less than $|0.03|$) (see table 2). Only the retrospective likelihood method for clinic-

Table 2. Comparison of four genetic models in penetrance estimates at age 70 among carriers: Bias, robust standard error (SE) and root mean square error (RMSE) from various study designs, using different likelihood methods

Method	Design	Dominant model with rare gene (q = 2%)						Recessive model with common gene (q = 30%)					
		high penetrance risk = 63%			low penetrance risk = 37%			high penetrance risk = 63%			low penetrance risk = 37%		
		Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE
Prospective	POP	<0.01	0.09	0.09	-0.01	0.12	0.12	<0.01	0.07	0.07	<0.01	0.09	0.09
	POP+	<0.01	0.06	0.06	<0.01	0.06	0.06	<0.01	0.06	0.06	<0.01	0.07	0.07
	CLI	-0.01	0.10	0.10	<0.01	0.11	0.11	<0.01	0.07	0.07	<0.01	0.09	0.09
	CLI+	<0.01	0.09	0.09	<0.01	0.10	0.10	<0.01	0.07	0.07	<0.01	0.08	0.08
Retrospective	POP	0.01	0.09	0.09	0.01	0.10	0.10	-0.01	0.07	0.07	<0.01	0.08	0.08
	POP+	<0.01	0.06	0.06	<0.01	0.06	0.06	<0.01	0.06	0.06	<0.01	0.07	0.07
	CLI	0.05	0.13	0.14	0.02	0.17	0.17	-0.16	0.12	0.20	-0.13	0.13	0.18
	CLI+	0.06	0.13	0.14	0.02	0.16	0.16	-0.16	0.12	0.20	-0.11	0.12	0.16
Joint	POP	0.01	0.09	0.09	0.01	0.10	0.10	-0.01	0.07	0.07	<0.01	0.08	0.08
	POP+	<0.01	0.06	0.06	<0.01	0.06	0.06	<0.01	0.06	0.06	<0.01	0.07	0.07
	CLI	0.01	0.09	0.09	0.01	0.11	0.11	-0.03	0.07	0.07	-0.03	0.08	0.09
	CLI+	0.01	0.08	0.09	<0.01	0.10	0.10	-0.02	0.07	0.07	-0.01	0.08	0.08

Table 3. Summary of the effective choices of study designs and likelihood methods for estimating relative risk and penetrance function when no second gene variation is involved; based on RMSE values

Genetic model	Study design	Likelihood method
<i>Relative risk estimation</i>		
Dominant model with rare gene		
High/low penetrance	CLI	retrospective or joint
Recessive model with common gene		
High penetrance	POP/CLI	retrospective or joint
Low penetrance	CLI	retrospective
<i>Penetrance estimation</i>		
Dominant model with rare gene		
High/low penetrance	POP+	any of the three likelihoods*
Recessive model with common gene		
High/low penetrance	POP+	any of the three likelihoods*
* Three likelihoods have the same form under the study design POP+.		

based designs had substantial bias (|0.07–0.14|) under the two dominant and the high-penetrance recessive models. So, the penetrance value could have an impact on the bias for this latter setting. The SEs tended to be much smaller than the corresponding values for the log relative risk estimates. The most efficient design was always the population-based design with mutation carrier probands (POP+) and this was true regardless of the genetic model considered. The SEs varied between 0.06 and 0.07 and had very similar values under all three likelihood methods. The level of penetrance had no noticeable impact on the SEs. Table 2 also confirms that the three likelihood methods were equivalent under POP+ design to estimate penetrance. The differences in SEs were also reflected in the RMSEs as the absolute biases were small compared to the SEs.

Key results about the design efficiency and the estimating methods in the absence of a second gene effect are summarized in table 3.

Simulations with Second Gene Variation

We describe in this section the results about the high-penetrance dominant model, but these results generalize to the other genetic models investigated.

Log Relative Risks

When a second gene effect was absent or relatively small (RR = 2.0) in the family data, the estimates in table 4 appeared almost unbiased across the three different likelihood methods. This was not true anymore when a large second gene effect (RR =

Table 4. Bias, robust standard error (SE), root mean square error (RMSE), and asymptotic relative efficiency (ARE) relative to the joint likelihood for log relative risk estimates of the major gene effect from various study designs, using different likelihood methods in the dominant model with a rare but highly penetrant allele

Method	Design	Second gene variation											
		none (RR = 1)				small (RR = 2)				large (RR = 5)			
		Bias	SE	RMSE	ARE	Bias	SE	RMSE	ARE	Bias	SE	RMSE	ARE
Prospective	POP	0.03	0.35	0.35	0.51	-0.01	0.31	0.31	0.60	-0.26	0.28	0.38	0.62
	POP+	0.04	0.34	0.34	1.00	-0.02	0.29	0.30	1.00	-0.26	0.23	0.35	1.00
	CLI	0.04	0.32	0.33	0.19	<0.01	0.30	0.30	0.22	-0.12	0.26	0.29	0.29
	CLI+	0.07	0.32	0.33	0.56	0.02	0.29	0.29	0.58	-0.15	0.25	0.29	0.64
Retrospective	POP	0.07	0.25	0.26	1.00	0.03	0.24	0.24	1.00	-0.18	0.22	0.29	1.00
	POP+	0.04	0.34	0.34	1.00	-0.02	0.29	0.30	1.00	-0.26	0.23	0.35	1.00
	CLI	0.05	0.17	0.17	0.68	0.02	0.16	0.16	1.00	-0.10	0.14	0.18	1.00
	CLI+	0.01	0.24	0.24	1.00	-0.03	0.23	0.23	0.77	-0.19	0.21	0.28	0.91
Joint	POP	0.07	0.25	0.26	-	0.03	0.24	0.24	-	-0.18	0.22	0.29	-
	POP+	0.04	0.34	0.34	-	-0.02	0.29	0.30	-	-0.26	0.23	0.35	-
	CLI	0.05	0.14	0.15	-	0.02	0.14	0.14	-	-0.09	0.14	0.17	-
	CLI+	0.04	0.24	0.24	-	0.01	0.22	0.22	-	-0.14	0.20	0.25	-

Table 5. Bias, robust standard error (SE), root mean square error (RMSE), and asymptotic relative efficiency (ARE) relative to the joint likelihood for penetrance estimates at age 70 among carriers from various study designs, using different likelihood methods in the dominant model with a rare but highly penetrant allele

Method	Design	Second gene variation											
		none (RR = 1)				small (RR = 2)				large (RR = 5)			
		Bias	SE	RMSE	ARE	Bias	SE	RMSE	ARE	Bias	SE	RMSE	ARE
Prospective	POP	<0.01	0.09	0.09	1.00	0.01	0.08	0.08	1.00	0.08	0.06	0.10	1.00
	POP+	<0.01	0.06	0.06	1.00	0.01	0.05	0.05	1.00	0.07	0.04	0.08	1.00
	CLI	-0.01	0.10	0.10	0.81	0.03	0.08	0.08	0.77	0.11	0.05	0.12	0.64
	CLI+	<0.01	0.09	0.09	0.79	0.03	0.08	0.08	0.77	0.10	0.05	0.11	0.64
Retrospective	POP	0.01	0.09	0.09	1.00	0.02	0.08	0.08	1.00	0.09	0.06	0.11	1.00
	POP+	<0.01	0.06	0.06	1.00	0.01	0.05	0.05	1.00	0.07	0.04	0.08	1.00
	CLI	0.05	0.13	0.14	0.48	0.06	0.11	0.12	0.41	0.11	0.07	0.13	0.33
	CLI+	0.06	0.13	0.14	0.38	0.07	0.10	0.12	0.49	0.11	0.06	0.13	0.44
Joint	POP	0.01	0.09	0.09	-	0.02	0.08	0.08	-	0.09	0.06	0.11	-
	POP+	<0.01	0.06	0.06	-	0.01	0.05	0.05	-	0.07	0.04	0.08	-
	CLI	0.01	0.09	0.09	-	0.04	0.07	0.08	-	0.12	0.04	0.13	-
	CLI+	0.01	0.08	0.09	-	0.04	0.07	0.08	-	0.11	0.04	0.12	-

5.0) was present and causing substantial residual familial correlation. In this case, the estimates of β_1 clearly appeared to be negatively biased for any choice of study design or likelihood methods. Thus, the log relative risk for the major gene would be substantially underestimated. The presence of a second gene effect did not affect the relative efficiency of the different designs. Using the retrospective and joint likelihood methods, the study designs with randomly selected affected probands (CLI) still provided

more precise estimates of β_1 . In the presence of a second gene, the standard errors of the log relative risk estimate decreased as the second gene variation increased. Figure 2 depicts the point and 95% confidence intervals (CIs) of the bias of the log relative risk estimate across the likelihood methods for different levels of second gene variation. The increasing trend in the absolute value of the bias as the second gene effect increased was evident across all three likelihood methods and the 95% CI for the bias was also

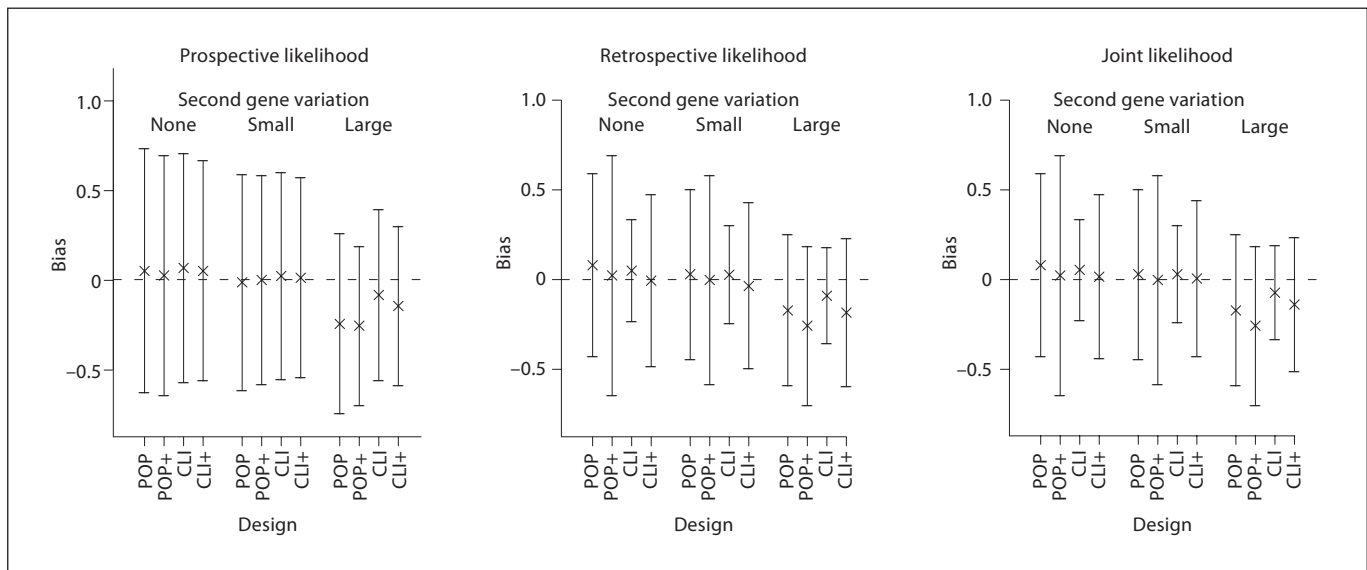


Fig. 2. The accuracy and precision in the log relative risk estimation of the major gene; the point and 95% confidence interval estimates of the bias based on the simulation study.

shorter for the CLI design for the retrospective and joint likelihood methods. Table 4 shows similar patterns in the RMSE in the presence of second gene variation. The RMSE values for the study design CLI still had the most combined efficient and unbiased log relative risk estimates for the major gene effect in the presence of second gene variation.

Penetrance Estimates

As observed for the RR estimates, the bias in the penetrance estimates was almost negligible with no or small second gene variation but could be substantial when there was a strong second gene effect (table 5). The bias was in the opposite direction to the bias in log relative risk estimates. Thus, penetrance tended to be overestimated when there was a strong effect from a second gene which is shared within families. The study design POP+ yielded the most precise penetrance estimates among carriers for any amount of disease risk heterogeneity for this genetic model. In addition, the presence of second gene variation affected the standard error of the penetrance estimate; the standard error decreased as the effect of the second gene increased. Overall, the population-based design with mutation-carrier probands (POP+) had the lowest RMSE for estimating penetrance among carriers regardless of the likelihood methods or the presence of residual familial correlation. Table 5 also indicates that the joint likelihood was always efficient for estimating lifetime disease risk compared to the prospective and retrospective likelihoods for all levels of the second gene variation. The prospective likelihood method was once again proportional to the joint likelihood for the POP+ design and the retrospective likelihood for both the POP and POP+ designs. Figure 3 depicts the point and 95% CIs of the bias of the penetrance estimate by sex and mutation carrier status for differing levels of second gene variation. The increasing trend in the bias as the second gene effect increased was evident across all four settings, but especially so among the carriers.

Application

We applied the three proposed likelihood methods to a real data set comprised of HNPCC pedigrees from Newfoundland who share a founder mutation in the MSH2 gene [13]. This data set of 343 phenotyped individuals was distributed among 12 very large families identified from a high risk criterion. Each family had a carrier proband, which corresponded to the study design CLI+. We adopted a Weibull hazard function to model the dependence of age at onset of colorectal cancer on MSH2 carrier status and gender. Since colorectal cancer constitutes a large component of HNPCC-related cancers, this phenotype was the primary outcome of interest. We estimated the penetrance using the retrospective likelihood method that we implemented in the genetic software MENDEL [14]. The retrospective likelihood was chosen because of the complex ascertainment used to identify these families, though either the prospective or joint likelihood methods would be more efficient if the ascertainment process could be modeled. The analysis of the Newfoundland data indicated that the relative risk of the MSH2 gene mutation effect on the age at onset of colorectal cancer was $e^{2.45} = 11.59$ and the penetrance of colorectal cancer by age 70 years was 91% for male carriers and 40% for female carriers. These values are comparable with those obtained using Kaplan-Meier survival curves (92% for male carriers and 64% for female carriers) in Green et al. [13], although the data sets are not identical. We also estimated the relative risk associated with an unobserved second gene to be $e^{1.01} = 3.0$. These results were close to the small second gene variation in our simulation study, so we would expect little impact from residual familial correlation in these data.

Based on these results, we then studied the relative performance of the designs POP, POP+ and CLI for risk estimation to the CLI+ design using the same level of precision. The asymptotic relative efficiencies (AREs) of study designs POP, POP+, CLI relative to CLI+ for estimating the log relative risks in the pres-

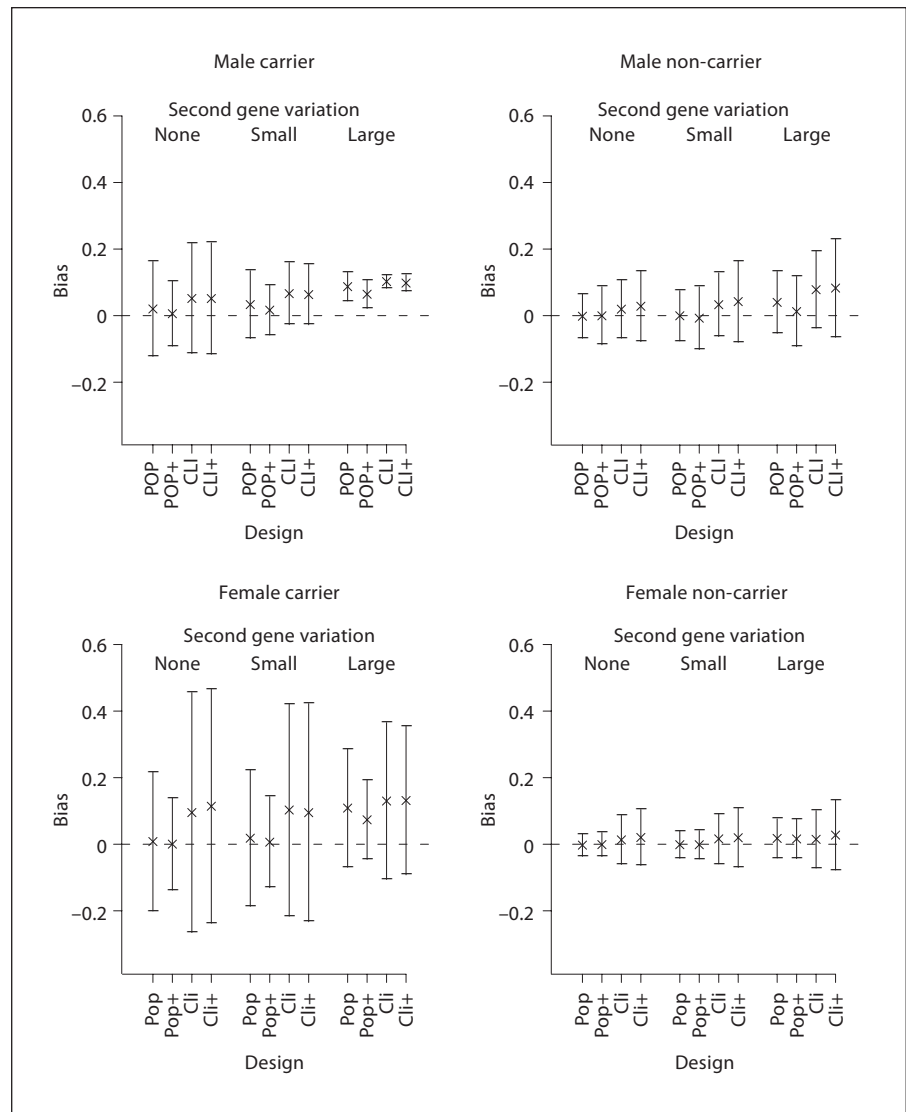


Fig. 3. The accuracy and precision in the penetrance estimation at age 70 using the joint likelihood method under the dominant model with a rare but highly penetrant disease gene; the point and 95% confidence interval estimates of the bias based on the simulation study.

ence of major and small second gene effects were found to be 0.96, 0.64 and 2.83, respectively. The AREs for estimating absolute risks were found to be 2.60, 5.82 and 0.97, respectively. Since the AREs can be interpreted as the ratio of the sample sizes needed for two study designs to yield the same efficiency, we calculated the number of families needed in the three other study designs in order to achieve the same efficiency as in the Newfoundland study using the CLI+ design. To estimate the log relative risk of the MSH2 gene mutation with an average number of 12 phenotyped individuals in each family, we would need 30, 44 and 10 families for the designs POP, POP+ and CLI, respectively. For estimating the penetrance at 70 years, the study designs POP, POP+ and CLI would require the collection of about 11, 5 and 29 families, respectively.

In practice, we may collect smaller population-based families, requiring more families to reach the same precision. Assuming that the average sample size of the two population-based designs

was smaller than the clinic-based designs, say 8 instead of 12 phenotyped individuals per family, then the number of families required for estimating the log relative risk and the penetrance by age 70 increases about 1.5 times the number needed for larger families. The numbers needed for estimating the log relative risk and the penetrance by age 70 using design POP are now 45 and 17, respectively, and 66 and 8 families using the design POP+. We should note that the penetrance estimates from the 12 pedigrees in Newfoundland appeared relatively imprecise with large confidence intervals by age 70 years, i.e. approximate 95% CI = (0.43, 1) for male carriers and (0.06, 0.99) for female carriers. Therefore, more accurate penetrance estimates among the gene carriers would be obtained by choosing a more efficient design (POP+) and by increasing markedly the number of pedigrees. For example, collecting four times more families is expected to reduce the SE of the penetrance estimate by half.

Discussion

A critical problem in gene characterization study is the choice of an appropriate design for reliable estimation of disease risk. While several family-based designs have been proposed, essentially either population- or clinic-based, their relative advantage remains unclear (Thomas [15] for a review). In this article, we considered four commonly-used family-based designs – population- and clinic-based with affected or affected mutation carrier probands – and we evaluated their performance in terms of both accuracy and efficiency under four genetic models of dominant and recessive models with high or low penetrance.

Our main conclusion is that the population-based design with affected mutation carrier probands (POP+) provides more efficient penetrance estimation and this result holds regardless of the genetic model investigated or the likelihood method considered. This is a very important result because the POP+ design is a simple study design and the correction for ascertainment can easily be performed under all likelihood formulations. In more practical situations, it might be difficult to collect families with a mutation-carrier proband. However, with the emergence of large international consortiums such as the NCI funded Breast and Colon Cancer Familial Registries (CFR) (<http://www.cfr.epi.uci.edu/>), the planning of studies using designs POP+ and CLI+ is now quite feasible. For example, as of September 2003, the Familial Breast Cancer Registry had identified 6,126 population-based case families and 1,647 clinic-based families with an affected proband [16]. They have also identified 230 population-based female mutation carriers for BRCA1 or BRCA2 genes, which is far higher than the sample size of 100 probands considered in our simulation studies. From 1998 to December 2005, the Colon CFR enrolled a total of 6,731 case families and 1,028 families from clinic-based sites (John McLaughlin, personal communication). The Colon CFR has also identified approximately 164 probands with a germline mutation in a MMR (mismatch repair) gene. The Colon CFR has family information and DNA from the relatives of these individuals, which can be used to estimate the penetrance for colorectal and other cancers. Therefore, the use of 100 families in the POP+ design, as specified in our simulation study, seems to be a reasonable sample size and the efficiency gains with more families in a study would be greater.

The second main conclusion concerning the study design issue is that the clinic-based design which does not select specifically mutation carrier probands (CLI), was

the most efficient for estimating the relative risk. The analysis of such designs, in general, requires more complex corrections for ascertainment so the use of an ascertainment-free likelihood method, i.e. the retrospective likelihood, would be appropriate but at the expense of some efficiency. Interestingly, we found that the loss of efficiency resulting from the use of the retrospective likelihood versus the joint likelihood was negligible in our simulation studies.

We therefore observed two fundamental results. First, a given design might be efficient for one question of interest but not for the other and, second, when considering the most efficient design for a particular question, an ascertainment correction seems possible using regular likelihood methods.

The study design employed for the 12 large HNPCC families collected in Newfoundland was of type CLI+. To estimate the log relative risk of the MSH2 gene mutation with an average family size of 12 phenotyped individuals, we would need 30, 44 and 10 families for the designs POP, POP+ and CLI, respectively. For estimating the penetrance at 70 years, the study designs POP, POP+ and CLI would require the collection of about 11, 5 and 29 families, respectively. Therefore, a more efficient design could have been used depending on the question of interest.

Our simulation study results contrast somewhat with those of Gong and Whittimore [1] who found that clinic-based designs (CLI) yielded more precise penetrance estimates than population-based designs (POP). Several possible reasons might explain these apparent discrepancies. First, these authors did not consider including mutation carrier probands in their designs. Second, their population-based design included two-stages and some loss of efficiency can result from the use of the Horvitz-Thompson estimator [8]. Finally, it should be noted that under the joint likelihood (which gives the most efficient penetrance estimates), the difference between the designs POP and CLI is very small in our simulations, and the gain in efficiency is mainly observed when considering mutation carrier probands only (design POP+). The increased efficiency of the clinic-based designs for estimating genetic relative risk and its decreased efficiency to estimate absolute penetrance is also noticed by Kraft and Thomas [3] for binary outcomes. This is because clinic-based designs do not provide much information on the baseline hazard, so are not very efficient for estimating absolute risk in non-carriers (and thus in carriers because the risk estimates in carriers and non-carriers are correlated). An appealing alternative to bolster the precision of absolute risk estimates from clinic-based designs is to

combine relative risks from the clinic-based design with population incidence rates (if available). Another potential disadvantage of the analysis of clinic-based designs is that it could be more sensitive to the specification of the allele frequency. Indeed, we observed a substantial bias in the regression parameters in our simulations under a recessive model with a common disease allele frequency (or dominant model with allele frequency above 5%, results not shown) for the two clinic-based designs analyzed with retrospective or joint likelihoods. These results hold whether or not a second gene effect was considered in the simulation. In the contrary, the prospective likelihood is robust to a misspecification of the allele frequency but probably less appropriate for complex ascertainment.

Finally, although we studied these designs individually, an optimal strategy could involve sampling families from the four designs studied here, where the optimal proportion could depend on the aim of the study and the underlying genetic model [8, 9]. This strategy could also be adapted if there is a need to estimate both the relative and cumulative risks. The objective function to be minimized could then depend jointly on these two measures of risk.

The efficiency gains must be considered in light of the existence of potential biases and limitations when planning a new gene characterization study. In our simulation study, we found that the presence of residual familial correlation due to a second gene can bias risk estimates. There was a negative bias in the log relative risk (bias in the direction of the null value of 0) and an overestimation of the penetrance estimate. This latter result was also found by two other studies [2, 4]. Interestingly, we found that the bias due to the second gene was smaller compared to the standard error of the risk estimate (genetic relative risk or penetrance), except when the second gene effect was large or in the presence of $G \times G$ interaction with small second gene effect. Investigating residual familial correlation in penetrance studies is therefore important for interpreting a major gene effect [1].

Several potential limitations of our study are worth mentioning. First, we assumed that we have complete data in our simulation study (i.e., all individuals are genotyped and parents are alive until 90 years of age). Incomplete information might influence the efficiency of the estimation but the relative efficiency should not be affected as long as the missing process is random. Missing genotypes in family studies can be inferred from the distribution of observed genotypes within the family, assuming Mendelian transmission of the alleles studied. However, the presence of both missing genotypes and phenotypes could lead to more serious biases, especially

if the process of missingness is not random. For example, if gene carriers die young, they will not be available for study. Thus, the study population can yield biased estimates for the disease of interest, which itself may or may not be a lethal disease. Methods such as Bayesian estimation, for example BayesMendel [17] for penetrance studies, could help in that situation, but this was not investigated in this study.

Second, for complex disease, the disease etiology can involve several genes and their interactions. We performed additional simulation studies to investigate the bias assuming the presence of interaction between the major gene and the second gene ($G \times G$ interactions) under a dominant model with rare disease allele for the major gene (results not shown). For certain interaction models, the bias in the parameters could become larger than their standard errors. This provides additional support for the results obtained when the second gene effect is very strong but without an interaction effect.

Third, another source of potential bias in penetrance estimation is the non-random selection of case probands. Begg [4] suggested several approaches to reduce the potential selection bias, which can be useful for the methods we have developed. In addition, we are also considering jointly modeling the ascertainment and the familial correlation using frailty terms in our future work.

Forth, fitting a PH model using the correctly specified parametric distribution for the baseline hazard used in our simulation studies overlooks the issue of model misspecification. A piecewise constant approach might be a suitable compromise for estimating the baseline hazard function. A piecewise-constant hazard function, with a small number of pieces and suitable selection of cut points, can approximate the true underlying hazard without relying on a fully parametric model [18, 19]. Spline or piecewise polynomial functions could also be used to estimate the baseline hazard function. Regression splines and their order-1 derivatives have the advantage of being continuous. But they also require cut points or knot locations to be selected and can include more unknown parameters than a piecewise-constant model. We are investigating the impact of model misspecification and the benefits of a weakly parametric approach in the designs and ascertainment corrections developed in this current study.

Last, the design comparisons that we described in our simulation study assumed the same number of families was included in each design. In more practical situations, the comparison of the different designs could also take into consideration the cost of the recruitment process.

Despite of these limitations, this work provides a comprehensive methodological framework for analyzing family-based designs in gene characterization studies and provides more rationale for the choice of an efficient design and appropriate likelihood method to estimate the risk associated with an inherited gene mutation.

In future work, we are planning to incorporate the residual familial correlation, which is a common feature of family data, directly into the modeling approach. Several authors have considered this problem [2, 4, 20–22] and have developed methods which allow for residual familial correlation in the analysis [21, 22]. However, a general methodology that could be applied to various designs and to various patterns of residual correlations is still lacking. Building on our previous work [23], we intend to develop frailty models to capture family dependence caused by latent common risk factors such as unknown genes shared within a family in the context of various family designs.

Acknowledgments

We thank the reviewers for their constructive comments and suggestions. This research was supported by a grant from the Institutes of Genetics and Population and Public Health of the Canadian Institutes of Health Research (Grant # 110053), an Interdisciplinary Health Research Team award from the Canadian Institutes of Health Research (Grant # 43821) and a fellowship from the Canadian Breast Cancer Foundation – Ontario Chapter.

Appendix 1

Estimating Model-Based Standard Errors for Penetrance Estimation

Let $\theta = (\beta_s, \beta_1, \rho, \lambda)^\top$ be the parameters in the model. The maximum likelihood estimates (MLEs), $\hat{\theta} = (\hat{\beta}_s, \hat{\beta}_1, \hat{\rho}, \hat{\lambda})^\top$, were obtained. Consider

$$S(t; \theta) = \exp[-\{\lambda(t - 20)\}^\rho e^{\beta_s x_s + \beta_1 x_1}]$$

$$LH(t; \theta) = \log\{-\log S(t; \theta)\} = \rho \log\{\lambda(t - 20)\} + \beta_s x_s + \beta_1 x_1.$$

The penetrance estimates, $F(t; \theta) = 1 - S(t; \theta)$, can be obtained by replacing θ with $\hat{\theta}$.

To compute the asymptotic variance of the penetrance estimate at fixed age t , we first use the log-log transformation of the survivor function, $LH(t; \theta)$, to work with a linear expression of the parameters and obtain the estimated variance of $LH(t; \theta)$, using the Delta method:

$$\text{Var}\{LH(t; \theta)\} = D_\theta^\top(t) \Sigma D_\theta(t)$$

where $D_\theta(t)$ is the vector of partial derivatives of $LH(t; \theta)$ with respect to each parameter,

$$D_\theta(t) = \frac{\partial LH(t; \theta)}{\partial \theta} = (x_s, x_1, \log\{\lambda(t - 20)\}, \rho/\lambda)^\top,$$

and Σ is the robust variance-covariance matrix of the parameters θ (see Appendix 2). The corresponding estimated variance can be obtained by replacing θ with $\hat{\theta}$ and using the approximate Σ evaluated at $\hat{\theta}$. Then, the resulting variance of penetrance estimate at age t can be approximated using the estimated variance of the $LH(t; \theta)$ by the Delta method again, i.e.,

$$\text{Var}\{F(t; \hat{\theta})\} = \text{Var}\{S(t; \hat{\theta})\} = \{e^{LH(t; \hat{\theta})} - e^{-LH(t; \hat{\theta})}\}^2 \text{Var}\{LH(\hat{\theta}; t)\}.$$

Appendix 2

Robust Variance Estimator

For consistent variance estimation, the variance matrix of the parameters $\theta = (\beta_s, \beta_1, \rho, \lambda)^\top$ is obtained by the sandwich estimator

$$\text{Var}(\theta) = H^{-1} V H^{-1},$$

where H is the matrix of second derivatives of the log-likelihood function and the variance of the score vector, V .

More specifically, we write the ascertainment-corrected likelihood in equation (1) in terms of the likelihood contribution of each individual for family f corrected by their being ascertained, i.e.,

$$L(\theta) = \prod_{f=1}^n \prod_{i=1}^{n_f} \frac{N_{fi}}{A_f},$$

resulting in the log-likelihood function, ℓ ,

$$\ell(\theta) = \sum_{f=1}^n \sum_{i=1}^{n_f} \log N_{fi} - \sum_{f=1}^n \log A_f.$$

The first two derivatives of the log-likelihood are obtained by

$$\ell'(\theta) = \sum_{f=1}^n \sum_{i=1}^{n_f} \frac{\partial}{\partial \theta} \log N_{fi} - \sum_{f=1}^n \frac{\partial}{\partial \theta} \log A_f;$$

$$\ell''(\theta) = \sum_{f=1}^n \sum_{i=1}^{n_f} \frac{\partial^2}{\partial \theta^2} \log N_{fi} - \sum_{f=1}^n \frac{\partial^2}{\partial \theta^2} \log A_f.$$

respectively, where we define

$$H = \ell''(\theta);$$

$$V = \text{Var}[\ell'(\theta)]$$

$$= \sum_f \text{Var} \left[\sum_{i=1}^{n_f} \frac{\partial}{\partial \theta} \log N_{fi} - \frac{\partial}{\partial \theta} \log A_f \right]$$

$$= \sum_f E \left[\left\{ \sum_{i=1}^{n_f} \frac{\partial}{\partial \theta} \log N_{fi} - \frac{\partial}{\partial \theta} \log A_f \right\} \right. \\ \left. \left\{ \sum_{i=1}^{n_f} \frac{\partial}{\partial \theta} \log N_{fi} - \frac{\partial}{\partial \theta} \log A_f \right\}^\top \right].$$

The robust sandwich variance estimator of $\hat{\theta}$ is obtained by evaluating at $\hat{\theta}$

$$\text{Var}(\hat{\theta}) = \hat{H}^{-1} \hat{V} \hat{H}^{-1}$$

References

- 1 Gong G, Whittemore AS: Optimal designs for estimating penetrance of rare mutations of a disease-susceptibility gene. *Genet Epidemiol* 2003;24:173–180.
- 2 Gail MH, Benichou J, Carroll R: Designing studies to estimate the penetrance of an identified autosomal dominant mutation: cohort, case-control, and genotyped-proband design. *Genet Epidemiol* 1999;16:15–39.
- 3 Kraft P, Thomas DC: Bias and efficiency in family-based gene-characterization studies: conditional, prospective, retrospective, and joint likelihoods. *Am J Hum Genet* 2000;66:1119–1131.
- 4 Begg CB: On the use of familial aggregation in population-based case probands for calculating penetrance. *J Nat Cancer Inst* 2002;94:1221–1226.
- 5 Thomas DC: *Statistical Methods in Genetic Epidemiology*. New York, Oxford University Press, 2004.
- 6 Le Bihan C, Moutou C, Brugieres L, Feunteun J, Bonaïti-Pellié C: ARCAD: a method for estimating age-dependent disease risk associated with mutation carrier status from family data. *Genet Epidemiol* 1995;12:13–25.
- 7 Carayol J, Bonaïti-Pellié C: Estimating penetrance from family data using a retrospective likelihood when ascertainment depends on genotype and age of onset. *Genet Epidemiol* 2004;27:109–117.
- 8 Whittemore AS, Halpern J: Multi-stage sampling designs in genetic epidemiology. *Stat Med* 1997;16:153–167.
- 9 Siegmund KD, Whittemore AS, Thomas DC: Multi-stage sampling for disease family registries. *J Nat Cancer Inst Monogr* 1999;26:43–48.
- 10 Iversen ES, Chen S: Population-calibrated gene characterization: estimating age at onset distributions for cancer genes. *J Am Stat Assoc* 2005;100:399–409.
- 11 Ewens WJ, Shute NCE: A resolution of the ascertainment sampling problem. I: Theory. *Theo Pop Biol* 1986;30:388–412.
- 12 Hodge S: Conditioning on subsets of the data: Application to ascertainment and other genetic problems. *Am J Hum Genet* 1988;43:364–373.
- 13 Green J, O'Driscoll M, Barnes A, Maher ER, Bridge P, Shields K, Parfrey PS: Impact of gender and parent of origin on the phenotypic expression of hereditary nonpolyposis colorectal cancer in a large Newfoundland kindred with a common MSH2 mutation. *Dis Colon Rectum* 2002;45:1223–1232.
- 14 Lange K, Weeks D, Boehnke M: Programs for pedigree analysis: MENDEL, FISHER, AND dGENE. *Genet Epidemiol* 1998;5:471–472.
- 15 Thomas DC: Design of gene characterization studies: an overview. *J Natl Cancer Inst Monogr* 1999;26:17–23.
- 16 John EM, Hopper JL, Beck JC, Knight JA, Neuhausen SL, Senie RT, Ziogas A, Andrulis IL, Anton-Culver H, Boyd N, Buys SS, Daly MB, O'Malley FP, Santella RM, Southey MC, Venne VL, Venter DJ, West DW, Whittemore AS, Seminara D: The Breast Cancer Family Registry: an infrastructure for cooperative multinational, interdisciplinary and translational studies of the genetic epidemiology of breast cancer. *Breast Cancer Res* 2004;6:R375–R389.
- 17 Chen S, Wang W, Broman K, Katki HA, Parmigiani G: BayesMendel: An R environment for Mendelian risk prediction. Johns Hopkins University Dept. of Biostatistics Working Papers 2004; Paper 39. <http://www.bepress.com/jhubiostat/paper39>.
- 18 Kim JS, Proschan F: Piecewise Exponential Estimator of the Survivor Function. *IEEE Trans Reliability* 1991;40:134–139.
- 19 Kalbfleisch JD, Prentice RL: *The Statistical Analysis of Failure Time Data*. New York, John Wiley and Sons, 1980, p 37.
- 20 Gail MH, Pee D, Carroll R: Effects of violations of assumptions on likelihood methods for estimating the penetrance of an autosomal dominant mutation from kin-cohort studies. *J Stat Plan Inference* 2001;96:167–177.
- 21 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.
- 22 Antoniou AC, Pharoah PDP, McMullen G, Day NE, Ponder BAJ, Easton DF: Evidence for further breast cancer susceptibility genes in addition to BRCA1 and BRCA2 in a population based study. *Genet Epidemiol* 2001;21:1–18.
- 23 Choi YH, Matthews DE: Accelerated life regression modeling of dependent bivariate time-to-event data. *Can J Stat* 2005;33:449–464.