18

On Combining Family Data from Different Study Designs for Estimating Disease Risk Associated with Mutated Genes

Yun-Hee Choi
University of Western Ontario
Canada

1. Introduction

Genetic disorders caused primarily by abnormalities in genes or chromosomes are rare in the general population. The associated putative mutations that lead to a high risk of developing such diseases are even rarer. In order to study disease risks associated with mutated genes, families sampled under different study designs are commonly used in association studies. This is because family data recruited via affected individuals (probands) would be expected to contain more affected individuals and mutation carriers than families randomly sampled from a general population, thus leading to increased statistical efficiency in estimating the disease risk. The disease risk associated with a mutated gene can be measured on a relative or absolute scale. As the event we consider is disease with its age of onset, the relative risk can be measured as a ratio of two hazards of developing disease between mutation carriers and non-carriers, and the absolute risk as a function of age, i.e., the cumulative risk of developing disease by a given age, which is also termed penetrance.

Several family-based study designs have been used for estimating the disease risk associated with a gene mutation when onset varies with age. Gong & Whittemore (2003) discussed 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 who are randomly sampled from the disease population. The proband is usually genotyped to determine if s/he carries the disease risk gene and additional genotype and phenotype data can then be collected from other family members. A kin-cohort design described by Wacholder et al. (1998) is an example of the population-based design as families are sampled through a volunteer (either affected or unaffected) who agrees to be genotyped and provides the disease history of her or his first-degree relatives through a questionnaire. Not restricted to including the first degree relatives and genotyping only probands, a kin-cohort design can be easily extended to case-family studies to include more extended family members and their genotype information. Case-control family studies have been widely used to analyse the ages of onset of disease in relation to genetic risk (Li et al., 1998; Shih & Chatterjee, 2000; Hsu & Gorfine, 2006), where case families are recruited via population-based cases and their matching control families are randomly sampled from the population.

For clinic-based designs, on the other hand, families are ascertained into the study 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. Such families are often identified from high-risk disease clinics and provide substantial information to estimate the disease risk (for example, Kopciuk et al., 2009). Multistage designs (Whittemore & Halpern, 1997; Siegmund et al., 1999) provide an alternative way to efficiently recruit high risk families, often using disease family registries, where families are sampled from more informative groups via several stages. Studies based on these high-risk families can be effective for characterizing the prevalence and penetrance of mutated genes, but it is well known that without proper ascertainment corrections statistical inference would lead to biased estimations of population attributes such as allele frequency, disease risks, and penetrance of the mutated genes.

To allow population-based inference for estimating disease risks associated with mutated genes, family data can be analyzed using various likelihood-based methods (Thomas, 2004). In particular, ascertainment-corrected likelihood approaches have been developed by several authors (for example, Choi et al., 2008; Carayol & Bonaïti-Pellié, 2004; Kraft & Thomas, 2000; Le Bihan et al., 1995). Based on the survival approach, Le Bihan et al. (1995) formulated a prospective likelihood for modeling phenotypes as the age of onset and disease status given genotypes, and corrected the likelihood by the probability of families being ascertained for study. This approach is natural as it models phenotypes as a function of genotype and covariates, but the ascertainment scheme has to be clearly known and simple enough to make proper correction. On the other hand, the retrospective likelihood models genotypes conditioning on the phenotypes of all family members (Carayol & Bonaïti-Pellié, 2004; Kraft & Thomas, 2000; Schaid et al., 2010). Although this approach provides the most robust way to obtain consistent estimates of relative risk even with the ascertainment schemes that are imprecisely defined or complex, it encounters the computational burden of summing over possible genotypes of all family members and a decreased efficiency resulting from conditioning. Choi et al. (2008) adapted the retrospective likelihood conditioning only on phenotypes of individuals who were involved in the ascertainment criteria; for families sampled from the population-based designs, only probands were used to correct for the ascertainment, whereas for families from the clinic-based designs, the probands and their parents and sibs were used for ascertainment correction. Moreover, Schaid et al. (2010) accommodated the composite likelihood approach to obtaining the retrospective likelihood based on all possible pairs of individuals in families to reduce the computational burden.

The main objectives of this article are first, to examine the effects of misspecification of study designs when more appropriate study designs have been ignored or incorrectly specified in the analysis; second, to provide simple and easy to apply adjustment schemes for estimating disease risks by combining family data from different study designs; and third, to develop an Expectation-Maximization algorithm to infer missing genotypes in the estimation of disease risks. We start with describing ascertainment-corrected likelihood methods to take the study design into account and propose a likelihood-based approach to estimating the disease risks for combined family data collected under different study designs. The performance of these ascertainment-corrected likelihood methods is evaluated in terms of bias and efficiency. The effect of design misspecification is examined for estimating the disease risks associated with mutated genes. The bias and efficiency involved in estimating two disease risks are

compared when only probands for the families from the clinic-based study are adjusted for, and when the probands and other affected family members for the families from the population-based design are used for ascertainment correction. For the combined family data, the two design correction methods (population-based and clinic-based) are applied and compared respectively with our proposed combined likelihood method in terms of their accuracies and efficiencies for disease risk estimation.

This chapter includes the following sections. Section 2 introduces two family-based study designs—the population- and clinic-based study designs and their ascertainment-corrected likelihood methods for modeling ages at onset for family members in disease risk estimation. We propose a likelihood-based approach for the combined family data obtained from different study designs. In Section 3, an Expectation-Maximization (EM) algorithm is incorporated to account for the missing genotype information, where the missing genetic covariates are inferred from their conditional expectation given the observed genotypes and phenotypes of other family members. Moreover, a robust variance estimator is proposed to account for the dependence of individuals within families. Using simulation studies in Section 4, we examine the effects of study design misspecification for estimating the disease risks and investigate the properties of our proposed likelihood approach for combined families from different study designs. In Section 5, we illustrate our proposed approaches through an application to family data obtained from the combination of two studies of Lynch Syndrome—first, Newfoundland data from the clinic-based design and second, Ontario data based on the population-based design. Final remarks and possible extensions of this work will follow in Section 6.

2. Methods

2.1 Defining disease risks

For diseases caused by mutated genes, the phenotype of interest varies in age at onset, i.e., time to an event such as death or disease diagnosis. We denote the age at onset by T, the affection status at age of examination by δ. Then, the phenotype is given by $D = (T, \delta)$. Under the Cox's proportional hazards model, the hazard function for individual i conditional on mutation gene G and other risk factors X is assumed to take the form

$$h(t_i|g_i, \boldsymbol{x}_i) = h_o(t) \exp(\beta_g G + \beta_x X),$$

where $h_o(t)$ is a baseline hazard function and β_g and β_x are unknown regression parameters.

Based on this model, we consider two types of disease risk associated with a mutated gene—relative risk and absolute risk, the latter is also called penetrance.

(1) the relative risk in survival analysis is defined by the hazards ratio for an individual with a mutated gene compared to an individual free from the mutation, that is

$$\textit{Relative Risk} = \exp(\beta_g).$$

(2) the penetrance function for the disease susceptibility gene is defined as the age-specific cumulative risk function conditional on the disease susceptibility gene G and other relevant covariates X,

$$\textit{Penetrance} = P(T < t|G, X)\,.$$

Design	Ascertainment Criteria
POP	Proband is affected
POP+	Proband is affected and 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

Table 1. Family-based study designs

The disease risks can be estimated by maximizing a likelihood function with proper ascertainment adjustment of the families. In a crude analysis of family data, their ascertainments are often corrected by simply excluding the probands from the analysis to prevent overestimating the risk. However, more prudent approaches such as likelihood methods would not simply drop out the probands because they include other important information about the disease risks. Rather they would adjust for the sampling process, allowing their contributions to the likelihood. To accommodate study designs and the ascertainment process properly, both the design and the ascertainment criteria should be known clearly. However, such designs or criteria in many cases are unclear or too complex to allow adjustment at the analysis stage. Moreover, family data could come from different sources where families were recruited using different designs or ascertainment criteria.

2.2 Family-based study designs

We consider the two most commonly used family-based study designs—population-based designs and clinic-based designs. The population-based study design uses the affected cases (probands) to sample their families while the clinic-based study design is based on the probands with a high family history of disease risk. Thus, the clinic-based families likely include more disease cases and mutation carriers compared to the families from population-based designs.

The ascertainment criteria for the population-based study are based on the affected probands who are randomly sampled from the diseased population; for example, cancer registries. To increase the power to study the effect of the mutated gene of interest, one can apply stringent criteria to recruit the probands to be not only affected but also be a mutation-carrier. Similarly, the clinic-based study designs can have two variants: one with random probands with multiple case family members and the other with carrier probands with multiple case family members. Such families can be recruited from cancer registries or cancer clinics.

Table 1 summarizes the four study designs and their sampling criteria used to ascertain families. Population-based designs correspond to ascertainment criteria POP and POP+. They are similar to a kin-cohort design but are more like case-family designs that include extended family members and their genotype information. 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.

2.3 Likelihood approaches for family-based study designs

This section describes the likelihood-based approaches for modeling ages at onset and genetic covariates using family data via population-based and clinic-based study designs. We propose a combined likelihood approach for family data arising from the two different study designs.

2.3.1 Ascertainment-corrected retrospective likelihood

The retrospective likelihood corrects for the ascertainment by conditioning on the phenotypes. Define $D = (d_1, \ldots, d_n)$ as a vector of phenotypes, $G = (g_1, \ldots, g_n)$ as a vector of genotypes, $X = (x_1, \ldots, x_n)$ a vector of covariates other than genotypes, and A the ascertainment event. The likelihood contribution L_f for a single family f can be written as

$$\begin{aligned} L_f &= P(G_f|D_f, X_f, A_f) \\ &= \frac{P(A_f|D_f, X_f, G_f)P(D_f|X_f, G_f)P(G_f)}{P(D_f, A_f|X_f)} \\ &\propto \frac{P(D_f|X_f, G_f)P(G_f)}{P(D_f, A_f|X_f)}\,, \end{aligned} \tag{1}$$

where we assume that $P(A_f|D_f, X_f, G_f)$ is equal to 1 if the vector D_f qualifies for ascertainment, and 0 otherwise, and so is independent of the parameter of interest.

We further assume that individuals' phenotypes are independent conditionally given their genotypes and covariates. Thus, we can express the numerator as

$$P(D_f|X_f, G_f) = \prod_{i=1}^{n_f} P(d_i|x_i, g_i)\,,$$

and

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

Here $P(g_i)$ is based on Hardy-Weinberg Equilibrium (HWE) and depends on the population allele frequency, $P(g_i|g_{m_i}, g_{f_i})$ is the Mendelian transmission probability given parents' genotypes (g_{m_i}, g_{f_i}) of individual i.

The denominator is the correction term used to account for the study designs. In the population-based study, the ascertainment correction is based on the proband's phenotype in that it equals the probability of the proband, p, being affected before his/her age at examination, a_p, i.e.,

$$P(D_f, A_f|X_f) = \sum_g P(T_p < a_p|g)P(g),$$

where the sum is over all possible genotypes of the proband. For POP+ design, the sum takes place by assuming the proband is a mutation carrier.

In the clinic-based study, the denominator is based on the phenotypes of four individuals, two parents and two sibs, who involved in their family's ascertainment process. It can be

expressed as

$$P(D_f, A_f|X_f) = \sum_{G_\omega} P(T_f < a_f|x_f, g_{\omega_f})^{\delta_f} P(T_f \geq a_f|x_f, g_{\omega_f})^{1-\delta_f} \times$$
$$P(T_m < a_m|x_m, g_{\omega_m})^{\delta_m} P(T_m \geq a_m|x_m, g_{\omega_m})^{1-\delta_m} P(g_{\omega_f}, g_{\omega_m}|g_{\omega_p}) \times$$
$$P(T_s < a_s|x_s, g_{\omega_s}) P(g_{\omega_s}|g_{\omega_p}) P(T_p < a_p|x_p, g_{\omega_p}) P(g_{\omega_p}),$$

where indices f, m, s, p represent father, mother, sib and proband, respectively, δ indicates the affection status, and $G_\omega = (g_{\omega_f}, g_{\omega_m}, g_{\omega_p}, g_{\omega_s})$ includes all possible genotypes of the four individuals in the ascertainment set. For CLI+ design, the sum in the denominator is taken over all possible genotypes, provided that the proband carries a mutated allele of the major gene. The conditional probabilities $P(g_{\omega_f}, g_{\omega_m}|g_{\omega_p} = 1)$ and $P(g_{\omega_s}|g_{\omega_p} = 1)$ are obtained based on the HWE and Mendelian transmission probabilities using Bayes theorem.

2.3.2 Combined population- and and clinic-based study designs

Consider a study where the families are sampled via different study designs, say n_p families from a population-based design and n_c families from a clinic-based design. When their study designs are known, we can construct the likelihoods based on their study designs. Let L_p and L_c be the likelihood functions based on the population-based design and clinic-based design, respectively. We propose the combined likelihood for the families from the two designs as

$$L_{comb}(\theta|D, G, X) = L_p(\theta|D^p, G^p, X^p) L_c(\theta|D^c, G^c, X^c),$$

where the superscripts p and c denote the population- and clinic-based study designs, respectively, and the likelihoods L_p and L_c are obtained using the retrospective likelihood approach in expression (1). Therefore, the combined likelihood using the retrospective likelihood approach can accommodate both population- and clinic-based designs.

Even when the sampling schemes are not clearly defined, we can still employ this combined likelihood approach by dividing the families into two groups—high risk and low risk families, according to the number of cases observed in the family. For example, a family would be classified as a high risk family if it includes at least three cases among family members, otherwise it would be classified as a low risk family.

3. Missing genotypes

In practice, family data often include some missing information, particularly, missing genotypes. In the presence of missing genotype information, we estimate the disease risks associated with a known gene mutation in the families. Suppose data include genetic covariates that consist of observed genotypes and missing genotypes and phenotypes as time-of-onset responses with no missing. To infer the unobserved genotypes in the family, we implement an expectation-maximazation (EM) algorithm (Dempster et al., 1977) and estimate the parameters in the likelihood. The EM algorithm is an iterative procedure that computes the maximum likelihood estimates (MLEs) in the presence of missing data.

3.1 Expectation-maximazation algorithm

Suppose a genetic covariate G_f in family f consists of observed genotypes G_{fo} and missing genotypes G_{fm} and the vector of unknown parameters θ includes both regression parameters

and baseline hazard parameters. In our situation, the expectation of the complete data (D_f, X_f, G_f), $f = 1, \ldots, n$, is taken with respect to the conditional distribution of missing genotypes G_{fm} given observed data (D_f, X_f, G_{fo}) and current estimates of θ. Then the parameter estimates are updated by maximizing the likelihood function using the estimate of missing data in the expectation step. These two steps iterate until convergence to obtain the MLEs, where the algorithm is guaranteed to increase the likelihood at each iteration.

The conditional expectation of the log-likelihood function $\ell(\theta|D, G, X)$ of the complete data (D, G, X) given the observed phenotypes D_o and genotypes G_o, or Q function for the k^{th} iteration is given by:

$$Q(\theta|\theta^{(k)}) = E_{\theta^{(k)}}\left[\ell(\theta|D, G, X)|D_o, G_o\right] . \tag{2}$$

For the i^{th} individual in family f, we can then obtain the conditional expectation of their missing genotype G_i given their observed phenotype D_i, covariates X_i, and the observed mutation status G_o of other family members, especially if the proband's genotype G_p is conditioned as:

$$\begin{aligned} &E_{\theta^{(k)}}[G_i|D_i, X_i, G_p = 1] = P_{\theta^{(k)}}(G_i|D_i, X_i, G_p = 1) \\ &= \frac{P_{\theta^{(k)}}(D_i|X_i, G_i)P(G_i|G_p = 1)}{P_{\theta^{(k)}}(D_i|X_i, G_i = 1)P(G_i = 1|G_p = 1) + P_{\theta^{(k)}}(D_i|X_i, G_i = 0)P(G_i = 0|G_p = 1)} . \end{aligned}$$

Here $P(G_i|G_p = 1)$ is the conditional probability of the mutation carrier status for family member i, using the family proband's known mutation status. Based on Mendelian transmission probabilities, we can express these as simple constants under an assumed genetic model. Under the model assumptions given above, the phenotype probabilities conditional on genotype status for the i^{th} individual can be expressed in terms of the hazard function h and the corresponding survival function S depending on his/her affection status δ_i as

$$P(D_i|X_i, G_i) = S(t_i; X_i, G_i)h(t_i; X_i, G_i)^{\delta_i} .$$

In the M step of the algorithm, we take the partial derivatives of Q with respect to θ and set to zero, that will maximize Q.

3.2 Robust variance estimator for the EM algorithm

We illustrate the use of robust variance estimators (sandwich estimators) to account for within-family dependencies for disease risk estimates. In the presence of missing genotypes, the variance estimators are modified accordingly upon the use of the EM algorithm (Louis, 1982).

Let $U(\theta)$ and $B(\theta)$ denote the score vector and the negative of the associated matrix of second derivatives for the complete data, respectively, and $U^*(\theta)$ and $B^*(\theta)$ be the corresponding vector and matrix for the incomplete data. Then, the observed information matrix can be expressed as

$$I_o(\theta) = E_\theta[B(\theta)|g_o, d_o] - E_\theta[U(\theta)U^\top(\theta)|g_o, d_o] + U^*(\theta)U^{*\top}(\theta), \tag{3}$$

where g_o and d_o denote the vectors of observed genotypes and phenotypes from data. At the maximum likelihood estimate of θ, because of the convergence of the EM algorithm, U^*

is zero. Thus, the observed information matrix can be obtained as the first two terms on the right hand side of (3) that arise from the complete data log-likelihood analysis. The first term is evaluated as

$$E_\theta[B(\theta)|g_o, d_o] = E_\theta\left[-\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta^\top}|g_o, d_o\right].$$

Oakes (1999) explicitly expressed the information matrix in terms of derivatives of the $Q(\theta|\theta^{(k)})$ function in equation (2) invoked by the EM algorithm, as given by

$$I_o(\theta) = \frac{\partial^2 \ell}{\partial\theta\partial\theta^\top} = \left\{\frac{\partial^2 Q(\theta|\theta^{(k)})}{\partial\theta\partial\theta^\top} + \frac{\partial^2 Q(\theta|\theta^{(k)})}{\partial\theta\partial\theta^{(k)\top}}\right\}_{\theta=\theta^{(k)}}, \tag{4}$$

where ℓ represents the observed data log-likelihood and the second term is viewed as the 'missing information.'

To account for familial correlation, as our model assumes the independence of the individuals in the family, we obtain the robust variance estimator for the ascertainment corrected likelihood with missing genotypes in a 'sandwich' form (White, 1982),

$$\mathrm{Var}(\hat{\theta}) = I_o(\theta)^{-1}\left\{\sum_{f=1}^{n} U_f^*(\theta)U_f^*(\theta)^\top\right\} I_o(\theta)^{-1}, \tag{5}$$

where $U_f^*(\theta)$ is the conditional expectation of the complete data score vector for family f given the observed data. Thus, the robust variance of $\hat{\theta}$ can be estimated by replacing θ by $\hat{\theta}$ in equation (5).

4. Simulation study

We carried out simulation studies to investigate the properties of our proposed likelihood methods and the effect of design misspecification. The simulation study aims to (1) assess bias and efficiency in disease risk estimation (relative risk and penetrance) for the retrospective likelihood-based approaches for family data from different study designs, (2) investigate potential bias and efficiency loss in risk estimation when the study designs are misspecified, and (3) evaluate the first two aims using combined data from two different study designs.

4.1 Family data generation

The simulation of family data is based on the method developed by Gauderman (1995) and further extended by Choi et al. (2008). We generated families of three generations: two parents and their two offspring, one of whom is the proband (affected individual from whom the family is selected). Each offspring has a spouse and their children ranged in number from two to five. At the first stage, all family members' ages at examination were obtained using a normal distribution with mean age 65 for the first generation and 45 for the second generation, with variance fixed at 2.5 years for both generations. It resulted in an average of 20 years difference between the parents and offspring. At the next stage, the proband's genotype of a major gene was determined conditioning on the proband's affection status by her/his age at examination, assuming Hardy-Weinberg equilibrium (HWE) with the fixed population allele frequency. Given the proband's genotypes, the genotypes of the other family members were then determined using HWE and Mendelian transmission probabilities calculated with Bayes'

formula. Once we simulated the age at examination and genotype information for all family members, then the time-to-onset of individual i was simulated from the proportional hazards model,

$$h(t_i|g_i) = h_0(t_i)\exp(\beta x_i),$$

where x_i indicates a carrier status of disease mutation gene for subject i and the baseline hazard is assumed to follow the Weibull distribution which has a form, $h_0(t) = \lambda\rho\{\lambda(t - 20)\}^{\rho-1}$.

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

$$T_p \sim T|T < a_p\,.$$

For the rest of family members, their times to onset were generated unconditionally. We also assumed the minimum age at onset was 20 years of age and the maximum age for followup was 90 years of age. Finally, the affection status δ_i for the ith individual was determined by comparing the age at onset T_i and age at examination a_i; $\delta_i = 1$ if $T_i < a_i$ and 0 otherwise.

4.2 Simulation study designs

Data were simulated under different configurations. We assumed Weibull baseline hazard functions with scale (λ) and shape (ρ) parameters equal to 0.01 and 3.2, respectively. This leads to a cumulative risk of 10% among mutation non-carriers by age 70. Two penetrances were considered: high and low penetrances corresponding to the log relative risk of a major gene (β) given by 2.4 and 1.8, respectively. The high penetrance represents a lifetime risk of 70% by age 70 among carriers of a major gene, which assumes a rare gene with the allele frequency 0.02 under the dominant model. The low penetrance provides a lifetime risk of 48% by age 70 among carriers.

We designed the simulation studies, first to investigate the effect of design misspecification, and second, to examine the properties of our proposed likelihood for combined family data from different study designs in the estimation of disease risks associated with a mutated gene.

(1) To study the effect of design misspecification, the study designs POP, POP+, CLI, and CLI+ were used to generate family data. For each design, two retrospective likelihood methods were applied to fit the data—one using correct adjustment of the study design and the other using a design with misspecified correction; for example, population-based ascertainment correction was used for the families under CLI+ design and clinic-based ascertainment correction was used for the families under POP+ design as for the misspecified design. We simulated 500 random samples of 200 families for each simulation configuration.

(2) To investigate potential bias and efficiency loss in disease risk estimation for the proposed likelihood approach for combined family data from population-based and clinic-based designs. We considered the combined families either from POP+ and CLI+ designs or POP and CLI designs with three mixing ratios between two designs—50-50, 70-30 and 80-20. For example, with the total 400 families sampled, the ratio 50-50 corresponds to equal numbers of families from POP+ and CLI+ designs, the 70-30 sampling corresponds to 280 POP+ families and 120 CLI+ families and the ratio 80-20 to 320 POP+ and 80 CLI+ families. The same numbers were examined for combining POP and CLI families. For each simulation configuration, 500 random samples were simulated.

4.3 Simulation results

Results of the simulation studies are described based on the empirical summary measures of bias and standard error obtained from the maximum likelihood estimates.

4.3.1 The effect of design misspecification

We first assessed bias and precision in disease risk estimation (relative risk and penetrance) for the retrospective likelihood with correct design adjustment for family data from different study designs. The results are summarized in Table 2.

With the correct design adjustment, the estimates of both the log relative risk and penetrance appeared unbiased; the absolute values of bias were less than 0.05 under both high and low penetrance models regardless of the study design. The magnitude of the bias was much smaller than the standard errors. In the log relative risk estimation, the precision of clinic-based designs was higher (smaller standard errors) than that of population-based designs. The population-based designs provided more accurate and precise estimates of the log relative risk for high penetrance than for low penetrance, whereas the clinic-based designs performed better for low penetrance. However, in the penetrance estimation, all designs provided more precise penetrance estimates (smaller standard errors) for high penetrance than for low penetrance.

We then examined the effect of design misspecification in terms of bias and precision of the log relative risk and penetrance estimates obtained from the retrospective likelihoods when the study design was misspecified. The clinic-based ascertainment correction was applied to the family data under the population-based designs and the population-based ascertainment correction to the clinic-based study. It is worth noting that the clinic-based design with the population-based correction provided relatively large bias in both disease risks, however, the bias in the population-based design with the clinic-based ascertainment correction was not notably large. Especially, under POP+ design (with affected and mutation carrier probands), the clinic-based retrospective likelihood yielded estimates at least as accurate as those from probands-only adjustment (correct design), although their standard errors were larger under the misspecified design.

4.3.2 The likelihood methods for combined family data from different study designs

We evaluated the accuracy and precision of the disease risk (log relative risk and penetrance) estimates based on the three retrospective likelihoods for combined data. Simulation results based on the combined data from CLI+ and POP+ designs are summarized in Table 3, and those from combining CLI and POP families in Table 4 .

Combined data from POP+ and CLI+ designs

In the log relative risk estimation, as expected, the population-based likelihoods for the combined data yielded overestimates because the ascertainment correction was based on only probands, which would not be sufficient for the families from clinic-based designs. However, the clinic-based retrospective likelihood provided slightly negative but less biased estimates in log relative risk but slightly larger standard errors. Although the population-based likelihoods provided smallest standard errors, they were subject to positive bias. Moreover, the log relative risk estimates for low penetrance performed better (less bias and higher precision) than for high penetrance. Our proposed likelihood was almost as efficient as the

Log relative risk (β) estimation

	High Penetrance ($\beta = 2.4$)				Low Penetrance ($\beta = 1.8$)			
	POP	POP+	CLI	CLI+	POP	POP+	CLI	CLI+
Correct Design	-0.002 (0.129)	0.010 (0.236)	0.044 (0.095)	-0.015 (0.171)	-0.006 (0.153)	0.017 (0.256)	0.017 (0.066)	-0.003 (0.136)
Misspecified Design	0.033 (0.159)	-0.022 (0.265)	1.456 (0.201)	0.444 (0.165)	0.041 (0.185)	0.009 (0.272)	0.665 (0.151)	0.475 (0.144)

Penetrance estimation

	High Penetrance (70%)				Low Penetrance (48%)			
	POP	POP+	CLI	CLI+	POP	POP+	CLI	CLI+
Correct Design	0.013 (0.049)	0.015 (0.033)	0.028 (0.078)	0.020 (0.084)	0.008 (0.057)	0.012 (0.040)	0.049 (0.115)	0.037 (0.103)
Misspecified Design	0.034 (0.087)	0.009 (0.098)	0.290 (0.003)	0.282 (0.004)	0.056 (0.140)	0.014 (0.135)	0.501 (0.003)	0.480 (0.006)

Table 2. Effects of the design misspecification: bias and precision in disease risk estimation based on retrospective likelihoods with correct and incorrect design adjustments; standard errors are in parenthesis.

population-based likelihood and as accurate as the clinic-based likelihood, regardless of the mixing rates we considered. Especially, the combined likelihood appeared to perform better for relative risk estimation when more CLI+ families were included in the sample.

In the penetrance estimation, we observed similar patterns as in the log relative risk estimation. The population-based likelihood provided substantially large bias with small standard errors, whereas the clinic-based likelihood yielded less bias with large standard errors. However, our proposed likelihood method offered the least bias and improved precision compared to the clinic-based likelihood. In addition, the penetrance was more precisely estimated with the combined likelihood when fewer CLI+ families were recruited (20% CLI+ families).

Combined data from POP and CLI designs

The patterns of bias and precision of the three likelihood methods were more clear with the combined data from POP and CLI designs, as shown in Table 4. In the log relative risk estimation, our proposed likelihood yielded both the most accurate and precise estimates. It also provided more precise estimates when 50% CLI families were included. Similarly, in penetrance estimation, the population-based likelihood provided heavily biased estimates; however, the combined likelihood performed well in terms of both bias and precision. With fewer CLI families (20%) in the data, more precise estimates were obtained.

Log relative risk (β) estimation

	High Penetrance ($\beta = 2.4$)			Low Penetrance ($\beta = 1.8$)		
POP+ vs. CLI+	50-50	70-30	80-20	50-50	70-30	80-20
POP+ corrected likelihood	0.279 (0.132)	0.196 (0.142)	0.145 (0.149)	0.326 (0.123)	0.240 (0.139)	0.191 (0.145)
CLI+ corrected likelihood	-0.024 (0.140)	-0.024 (0.154)	-0.026 (0.163)	-0.004 (0.124)	-0.010 (0.141)	-0.002 (0.150)
Combined likelihood	-0.025 (0.134)	-0.026 (0.143)	-0.028 (0.149)	-0.005 (0.123)	-0.011 (0.140)	-0.005 (0.147)

Penetrance estimation

	High Penetrance (70%)			Low Penetrance (48%)		
POP+ vs. CLI+	50-50	70-30	80-20	50-50	70-30	80-20
POP+ corrected likelihood	0.209 (0.009)	0.151 (0.015)	0.113 (0.017)	0.348 (0.012)	0.247 (0.017)	0.182 (0.020)
CLI+ corrected likelihood	0.019 (0.060)	0.016 (0.067)	0.015 (0.067)	0.032 (0.079)	0.021 (0.083)	0.024 (0.085)
Combined likelihood	-0.008 (0.031)	-0.011 (0.029)	-0.012 (0.027)	0.002 (0.033)	-0.002 (0.030)	-0.002 (0.028)

Table 3. Bias and precision in disease risk estimation based on three retrospective likelihood approaches for combined data from different family based designs (POP+ and CLI+) with affected and mutation carrier probands; standard errors are in parenthesis.

5. Application to Lynch Syndrome families

Lynch Syndrome, also referred to as hereditary non-polyposis colorectal cancer is an autosomal dominant condition which predisposes carriers to colorectal cancer (CRC). Several DNA mismatch repair (MMR) genes responsible for the majority of Lynch Syndrome cancers have been identified, predominantly MLH1 and MSH2. For the study of CRC, Lynch Syndrome families share a founder mutation in an MMR gene sampled from Newfoundland and Ontario. The Newfoundland data consist of 315 phenotyped individuals (74 affected and 241 not affected) from 12 very large families identified using a high risk criteria. Of them, 261 were genotyped (162 carriers, 99 non-carriers) and 54 were not genotyped. Each family had a carrier proband and other affected relatives, which corresponds to the study design CLI+. The Ontario data were identified through the Ontario Familial Colorectal Cancer Registry (Cotterchio et al., 2000) and consist of 506 phenotyped individuals (126 affected and 380 not affected) from 32 families with MMR mutation carrier probands, which corresponds to the

Log relative risk (β) estimation

	High Penetrance ($\beta = 2.4$)			Low Penetrance ($\beta = 1.8$)		
POP vs. CLI	50-50	70-30	80-20	50-50	70-30	80-20
POP corrected likelihood	0.911 (0.089)	0.644 (0.079)	0.485 (0.078)	0.700 (0.094)	0.609 (0.089)	0.506 (0.089)
CLI corrected likelihood	0.044 (0.079)	0.035 (0.086)	0.038 (0.094)	0.020 (0.063)	0.024 (0.077)	0.029 (0.087)
Combined likelihood	0.014 (0.072)	-0.009 (0.076)	-0.017 (0.080)	0.003 (0.058)	0.000 (0.068)	-0.002 (0.076)

Penetrance estimation

	High Penetrance (70%)			Low Penetrance (48%)		
POP vs. CLI	50-50	70-30	80-20	50-50	70-30	80-20
POP corrected likelihood	0.269 (0.005)	0.237 (0.010)	0.203 (0.013)	0.467 (0.007)	0.413 (0.012)	0.353 (0.018)
CLI corrected likelihood	0.043 (0.055)	0.041 (0.055)	0.042 (0.056)	0.059 (0.081)	0.060 (0.082)	0.062 (0.082)
Combined likelihood	0.006 (0.042)	-0.002 (0.038)	-0.005 (0.036)	0.015 (0.047)	0.008 (0.043)	0.006 (0.042)

Table 4. Bias and precision in disease risk estimation based on three retrospective likelihood approaches for combined data from different family based designs (POP and CLI) with random affected probands; standard errors are in parenthesis.

study design POP+. Of them, 154 individuals were genotyped (92 carriers, 62 non-carriers) and 352 were not genotyped.

The three likelihood methods (POP+ corrected, CLI+ corrected and combined likelihoods) were applied to combined families with Lynch Syndrome identified from Newfoundland (CLI+) and Ontario (POP+). A Weibull model was used to assess the effects of MMR mutation gene and gender on the age at onset of colorectal cancer. The EM algorithm was implemented to infer missing genotypes. The results of fitting these Lynch Syndrome families using different likelihood methods are presented in Table 5, and the age-specific penetrance estimates based on the combined likelihood are graphically illustrated in Figure 1.

In the analysis based on the combined likelihood, the β parameters for the genetic and gender effects were estimated to be 1.13 with robust standard error (se) = 0.18 and -0.51 with se=0.17, respectively, which lead to the hazards ratio of the MMR mutation carriers for the colorectal cancer as 3.10 (se=0.55) and the hazards ratio between female and male as 0.60 (se=0.11).

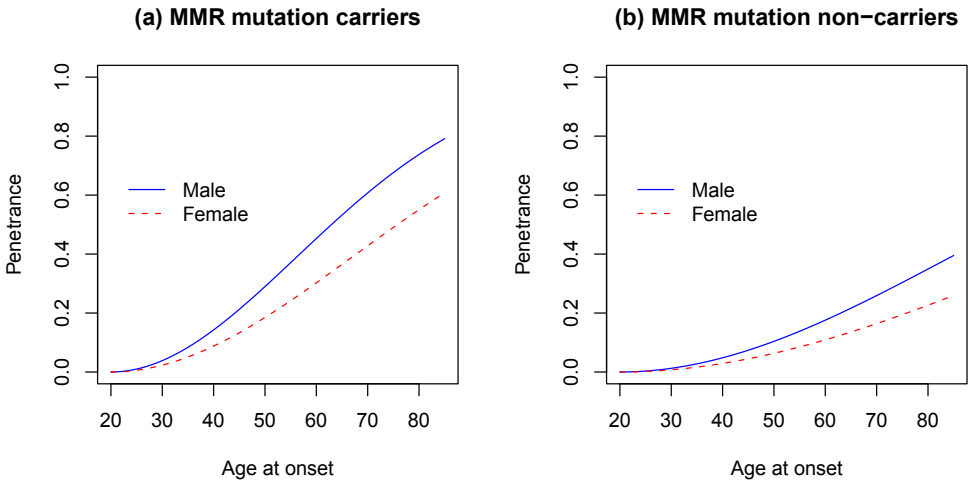

Fig. 1. (a) Estimated cumulative risk of developing colorectal cancer for carriers of any MMR gene mutation for the Lynch Syndrome families from Newfoundland and Ontario. (b) Same as (a) for non-carriers.

Log relative risk estimation in terms of hazards ratio

	MMR mutation		Gender	
POP+ corrected likelihood	1.07	(0.17)	-0.42	(0.16)
CLI+ corrected likelihood	1.15	(0.18)	-0.52	(0.18)
Combined likelihood	1.14	(0.18)	-0.51	(0.17)

Age-specific penetrance estimation among mutation carriers

	Male		Female	
POP+ corrected likelihood	62.4%	(4.08)	47.5%	(4.06)
CLI+ corrected likelihood	58.9%	(4.27)	40.9%	(4.18)
Combined likelihood	60.7%	(4.15)	42.8%	(4.10)

Table 5. Disease risk estimates and their corresponding robust standard errors in parenthesis using different likelihood methods for the Lynch Syndrome families from Newfoundland and Ontario

These relative risks indicated that the MMR mutation carriers were approximately three times more likely to develop the colorectal cancers than non-carriers, whereas among males and females, females showed about one third lower the hazard rate than males. There was very little difference observed between the relative risk estimates obtained by the CLI+ corrected

likelihood and the combined likelihood, although their precisions were slightly better with the combined likelihood.

We obtained that the penetrance of colorectal cancer by age 70 was 61% (se=4.15) among male carriers and 43% (se=4.1) among female carriers using the combined likelihood. These estimates were comparable with those obtained using the POP+ and CLI+ corrected retrospective likelihoods. Penetrances were overestimated (62% for male and 48% for female carriers) with higher precision (se=4.08 for male, 4.06 for female) under POP+ correction but slightly underestimated (59% for male carriers and 41% for female carriers) with lower precision (se=4.27 for male and 4.18 for female) under CLI+ correction, as seen in our simulation study.

6. Conclusion

In genetic epidemiology, family studies have been widely used for identifying genes responsible for traits and characterizing their risks in the population and they are often based on various family-based designs to sample families depending on the objectives of the study or their budget. To make population-based inferences, the study design should be properly taken into account, especially when the sampling is not randomly conducted as often is the case with the sampling of families.

In this study, for estimating disease risks—relative risk and penetrance, we have proposed the use of a retrospective likelihood to take the sampling process of families into account, and investigated the effect of sampling design misspecification on disease risk estimation. Our study showed that the misspecification of study design undoubtedly lead to bias; overestimation of risks when the study design adjustment was less than it should be (i.e, the clinic-based designs were analyzed with the correction by probands only), and underestimation with overcorrection by multiple affected family members. However, the magnitudes of bias and precision varied depending on the study design and the size of the penetrance. We found that undercorrection created more bias although it provided smaller standard error. This implies that conditioning more individuals would be safer for obtaining accurate estimates at the price of loss of precision if the study design is not known. The POP+ design with clinic-based correction in fact provided unbiased estimates of relative risk and penetrance. In general, the population-based designs performed better for high penetrance for estimating both disease risks but the clinic-based designs performed differently: penetrance was more efficiently estimated under high penetrance but relative risk was more efficiently estimated under low penetrance. In addition, we have proposed the combined likelihood for families sampled under different study designs and the effect of design misspecification was also investigated for combined data. Our proposed likelihood is applicable even when the study designs of the combined data are not clearly known since we can divide families into two categories—high risk families with at least three affected individuals and low risk families, otherwise. Our proposed combined retrospective likelihood method yielded accurate and precise estimates of both disease risks. Comparatively, the clinic-based likelihoods applied to combined data and provided unbiased estimates less efficiently compared to those from the combined likelihood. It is noteworthy that the EM algorithm we developed for inferring missing genotypes is a novel way to impute the missing genotypes using the observed genotypic and phenotypic information from other family members.

In practice, it might be difficult to collect families with a mutation-carrier proband. However, with the emergence of large international consortiums such as the Breast and Colon Cancer Family Registries, the planning of studies using designs POP+ and CLI+ is now quite feasible. Therefore, the use of 200 families in the CLI+ design, as specified in our simulation study, seems to provide a reasonable sample size; however, the efficiency gains with more families would clearly be greater.

There are potential limitations to our study. First, we assumed the Weibull distribution, chosen to model the penetrance function because of flexible modeling of the baseline hazard function which includes constant, increasing or decreasing hazard functions. There might be potential for model misspecification. Kopciuk et al. (2009) employed the generalized log-Burr model for more flexible modeling as it includes the Weibull model or the log-logistic model as special cases (Lawless, 2003), where the Weibull model has a monotonic functional form of the hazard whereas the log-logistic model does not. The baseline hazard can be also modeled semiparametrically using a step function while assuming proportional hazards. Second, between-family heterogeneity in allele frequencies and baseline hazards can lead to bias in parameter estimates based on the homogeneous models. A random effect model would allow us to take between-family heterogeneity into account while avoiding a great number of family-specific parameters. Finally, familial correlation is a common feature of family data due to the unobserved genetic or environmental risk factors shared within families. We did not explicitly model within-family dependencies, instead, we accommodated a robust variance estimator. However, ignoring familial correlation can lead to biased estimates of the model parameters, and so to biased disease risks (Choi et al., 2008). Relating to other work, several authors have adopted mixed effect models for binary outcomes in family studies (Heagerty, 1999; Pfeiffer et al., 2008; Zheng et al., 2010). Shared frailty models can allow us to model times to onset data from families while explicitly modeling familial correlation. We are planning to develop such frailty models in the context of various family designs.

7. Acknowledgment

This research was supported by the Canadian Institutes of Health Research-Interdisciplinary Health Research Team, Grant no. 43821, the Institutes of Genetics and Population and Public Health of the Canadian Institutes of Health Research, Grant no. 110053 and the Natural Sciences and Engineering Research Council of Canada.

8. References

Carayol, J. & Bonaïti-Pellié, C. (2007). Estimating Penetrance From Family Data Using a Retrospective Likelihood When Ascertainment Depends on Genotype and Age of Onset, *Genetic Epidemiology*, Vol. 27: 109–117, ISSN 1098-2272.

Choi, Y.-H.; Kopciuk, K.A. & Briollais, L. (2008). Estimating disease risk associated with mutated genes in family-based designs, *Human Heredity*, Vol. 66: 238–251, ISSN 0001-5652.

Cotterchio, M.; McKeown-Eyssen, G.; Sutherland, H.; Buchan, G.; Aronson, M.; Easson, A.M.; Macey, J.; Holowaty, E. & Gallinger, S. (2000). Ontario familial colon cancer registry: methods and first year response rates, *Chronic Diseases in Canada*, Vol. 21: 81–86, ISSN 0228-8699.

Dempster, A.P.; Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society. Series B*, Vol. 39:1–38, ISSN 1369-7412.

Gong, G. & Whittemore, A.S. (2003). Optimal designs for estimating penetrance of rare mutations of a disease-susceptibility gene, *Genetic Epidemiology*, Vol. 24:173–180, ISSN 1098-2272.

Green, J.; O'Driscoll, M.; Barnes, A.; Maher, E.R.; Bridge, P.; Shields, K. & Parfrey, P.S. (2002). 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, *Diseases of the Colon and Rectum*, Vol. 45:1223–1232, ISSN 1530-0358.

Heagerty, P.J. (1999). Marginally specified logistic-normal models for longitudinal binary data, *Biometrics*, Vol. 55: 688–698, ISSN 1541-0420.

Hsu, L. & Gorfine, M. (2006). Multivariate survival analysis for case-control family studies, *Biostatistics*, Vol. 7: 387–398, ISSN 1468-4357.

Kopciuk, K.A.; Choi, Y.-H.; Parkhomenko, E.; Parfrey, P.; McLaughlin, J.; Green, J. & Briollais, L. (2009) Penetrance of HNPCC-related cancers in a retrospective cohort of 12 large Newfoundland families carrying a MSH2 founder mutation: an evaluation using modified segregation models, *Hereditary Cancer in Clinical Practice*, Vol.7: 16, ISSN 1897-4287.

Kraft, P. & Thomas, D.C. (2000). Bias and efficiency in family-based gene-characterization studies: conditional, prospective, retrospective, and joint likelihoods, *The American Journal of Human Genetics*, Vol. 66: 1119–1131, ISSN 1537-6605.

Lawless, J.F. (2003). *Statistical Models and Methods for Lifetime Data*, (Second Ed.), John Wiley and Sons Inc., ISBN 9780471372158, Hoboken.

Le Bihan, C.; Moutou, C.; Brugières, L.; Feunteun, J. & Bonaïti-Pellié, C. (1995). ARCAD: a method for estimating age-dependent disease risk associated with mutation carrier status from family data, *Genetic Epidemiology*, Vol. 12: 13–25, ISSN 1098-2272.

Li, H.; Yang, P. & Schwartz, A.G. (1998). Analysis of age of onset data from case-control family studies, *Biometrics*, Vol. 54: 1030–1039, ISSN 1541-0420.

Oakes, D. (1999). Direct calculation of the information matrix via the EM algorithm, *Journal of Royal Statistical Society, Series B*, Vol. 61, 479–482, ISSN 1369-7412.

Pfeiffer, R. M., Pee, D. & Landi, M.T. (2008). On combining family and case-control studies, *Genetic Epidemiology*, Vol. 32:638–646, ISSN 1098-2272.

Schaid, D.J.; McDonnell, S.K.; Riska, S.M.; Carlson, E.E. & Thibodeau, S.N. (2010). Estimation of genotype relative risks from pedigree data by retrospective likelihoods, *Genetic Epidemiology*, Vol. 34:287–298, ISSN 1098-2272.

Shih, J.H. & Chatterjee, N. (2000). Analysis of survival data from case-control family studies, *Biometrics*, Vol. 58: 502–509, ISSN 1541-0420.

Siegmund, K.D.; Whittemore, A.S. & Thomas, D.C. (1999). Multistage sampling for disease family registries, *Journal of the National Cancer Institute Monographs*, Vol. 26: 43–48, ISSN 1745-6614.

Thomas, D.C. (2004). *Statistical Methods in Genetic Epidemiology*, Oxford University Press, ISBN-13 978-0195159394, New York.

Wacholder, S.; Hartge, P.; Struewing, J.P.; Pee, D.; McAdams, M.; Brody, L. & Tucker, M. (1998). The kin-cohort study for estimating penetrance, *American Journal of Epidemiology*, Vol. 148:623–630, ISSN 1476-6256.

White, H. (1982). Maximum likelihood estimation of misspecified models, *Econometrica*, Vol. 50:1–25, ISSN 1468-0262.

Whittemore, A. S. & Halpern, J. (1997). Multi-stage sampling designs in genetic epidemiology, *Statistics in Medicine*, Vol. 16: 153–167, ISSN 1097-0258.

Zheng, Y.; Heagerty, P.J.; Hsu, L. & Newcomb, P.A. (2010) On combining family-based and population-based case-control data in association studies, *Biometrics*, Vol. 66: 1024–1033, ISSN 1541-0420.

Epidemiology Insights

Edited by Dr. Maria De Lourdes Ribeiro De Souza Da Cunha

ISBN 978-953-51-0565-7
Hard cover, 396 pages
Publisher InTech
Published online 20, April, 2012
Published in print edition April, 2012

This book represents an overview on the diverse threads of epidemiological research, brings together the expertise and enthusiasm of an international panel of leading researchers to provide a state-of-the art overview of the field. Topics include the epidemiology of dermatomycoses and Candida spp. infections, the epidemiology molecular of methicillin-resistant Staphylococcus aureus (MRSA) isolated from humans and animals, the epidemiology of varied manifestations neuro-psychiatric, virology and epidemiology, epidemiology of wildlife tuberculosis, epidemiologic approaches to the study of microbial quality of milk and milk products, Cox proportional hazards model, epidemiology of lymphoid malignancy, epidemiology of primary immunodeficiency diseases and genetic epidemiology family-based. Written by experts from around the globe, this book is reading for clinicians, researchers and students, who intend to address these issues.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Yun-Hee Choi (2012). On Combining Family Data from Different Study Designs for Estimating Disease Risk Associated with Mutated Genes, Epidemiology Insights, Dr. Maria De Lourdes Ribeiro De Souza Da Cunha (Ed.), ISBN: 978-953-51-0565-7, InTech, Available from: http://www.intechopen.com/books/epidemiology-insights/on-combining-family-data-from-different-study-designs-for-estimating-disease-risk-associated-with-mu

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

© 2012 The Author(s). Licensee IntechOpen. This is an open access article distributed under the terms of the Creative Commons Attribution 3.0 License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.