Biometrics 72, 95–105
March 2016
DOI: 10.1111/biom.12380

# Optimum Study Design for Detecting Imprinting and Maternal Effects Based on Partial Likelihood

**Fangyuan Zhang,[1] Abbas Khalili,[2] and Shili Lin[1,*]**

[1]Department of Statistics, The Ohio State University, 1958 Neil Avenue, Columbus, Ohio 43210, U.S.A.
[2]Department of Mathematics and Statistics, McGill University, 805 Sherbrooke Street West, Montreal, Quebec H3A 0B9, Canada
*email:* shili@stat.osu.edu

Summary. Despite spectacular advances in molecular genomic technologies in the past two decades, resources available for genomic studies are still finite and limited, especially for family-based studies. Hence, it is important to consider an optimum study design to maximally utilize limited resources to increase statistical power in family-based studies. A particular question of interest is whether it is more profitable to genotype siblings of probands or to recruit more independent families. Numerous studies have attempted to address this study design issue for simultaneous detection of imprinting and maternal effects, two important epigenetic factors for studying complex diseases. The question is far from settled, however, mainly due to the fact that results and recommendations in the literature are based on anecdotal evidence from limited simulation studies rather than based on rigorous statistical analysis. In this article, we propose a systematic approach to study various designs based on a partial likelihood formulation. We derive the asymptotic properties and obtain formulas for computing the information contents of study designs being considered. Our results show that, for a common disease, recruiting additional siblings is beneficial because both affected and unaffected individuals will be included. However, if a disease is rare, then any additional siblings recruited are most likely to be unaffected, thus contributing little additional information; in such cases, additional families will be a better choice with a fixed amount of resources. Our work thus offers a practical strategy for investigators to select the optimum study design within a case-control family scheme before data collection.

Key words: Ascertainment; Association study; Imprinting effect; Maternal effect; Partial likelihood; Study design.

## 1. Introduction

Genomic imprinting and maternal effects are both important epigenetic factors that are involved in many complex human diseases, including Prader-Willi and Angelman syndromes (Lawson et al., 2013), and childhood cancers (Nousome et al., 2013). Genomic imprinting (maternal or paternal) is an effect of epigenetic process involving methylation and histone modifications in order to silence the expression of a gene inherited from a particular parent without altering the genetic sequence. Maternal effect, on the other hand, refers to a situation where the phenotype of an individual is influenced by the genotype of the mother regardless of one's own genotype. Though genomic imprinting and maternal effects arise from two different underlying epigenetic mechanisms, they can produce the same parent-of-origin patterns of phenotypic variation. As such, it is necessary to distinguish and study these two confounding effects together to avoid false positives and/or false negatives. There are a number of existing methods that do model imprinting and maternal effects simultaneously to avoid potential confounding. Such approaches include a **L**ikelihood inference method for detecting **I**mprinting and **M**aternal **E**ffects (LIME), which can utilize nuclear families with an arbitrary number of affected and unaffected children, no matter whether the father's genotype is missing or not (Yang and Lin, 2013; Han et al., 2013). LIME uses only part of the full likelihood—partial likelihood—by exploiting the fact that the part of the likelihood containing the parameters of interest can be separated from that containing the nuisance parameters. It thus alleviates the need to make typically unrealistic assumptions and thus leads to a robust procedure with potentially greater power.

Despite spectacular advances in molecular genomic technologies in the past two decades, resources available for genomic studies are still finite and limited, especially for family-based studies. Hence, it is important to consider an optimum study design to maximally utilize limited fixed resources to increase statistical power using LIME to detect imprinting and maternal effects simultaneously. The particular question of interest is whether it is more profitable to genotype siblings of probands (individuals through whom the families are recruited into the study) or it is more informative to recruit more independent families, keeping the total number of individuals needed to be genotyped fixed. Such a question is of great interest in genetic epidemiology in general, but the conclusions in the literature are mixed. There are studies showing that recruiting a smaller number of larger families is better than a larger number of smaller families (Zhou et al., 2009; Han et al., 2013), but there are also studies arguing for the reverse (He et al., 2011; Li et al., 2014). There are yet another set of articles that show both

may result depending on the underlying settings (Li and Cui, 2010; Sung and Rao, 2008). For LIME, in particular, Han et al. (2013) carried out a limited simulation study to investigate relative power for detecting association, imprinting, and maternal effects for several case-control family-based study designs having the same total number of individuals. They concluded that the results "suggest that collecting more siblings rather than more families is a more effective way to increase statistics power." However, the conclusion is far from settled as the evidence is weak and the conclusion is based on a limited simulation. This is also true for the other studies discussed above. That is, to date, there has been no rigorous statistical analysis to address the study design issue for detecting imprinting and maternal effects, to the best of our knowledge; rather, all results and recommendations in the literature are based on anecdotal evidence from limited simulation studies. Our work here is to try to fill this void.

In this article, we propose a systematic approach to study various study designs for simultaneous detection of imprinting and maternal effects based on a partial likelihood formulation. To enable such an investigation, we first derive the asymptotic properties of the partial likelihood method that we employ for simultaneous effect detection. In particular, we obtain closed-form formulas for computing the information contents, either family-based, or individual-based, of each study design that is being investigated. Our results show that the conclusion is more complex than any simple rule of thumb; rather, the conclusion is dependent on the prevalence of the disease.

## 2. Asymptotic Study and Information Calculation

### 2.1. *The LIME Procedure*

LIME considers a candidate marker with two alleles $M_1$ and $M_2$, where $M_1$ is the allele of interest, which may code for disease susceptibility or epigenetic effect. In a nuclear family, $F$ and $M$ are the genetic variables for father and mother, which is coded as 0, 1, or 2, corresponding to genotype $M_2M_2$, $M_1M_2$, or $M_1M_1$, respectively. For each child in the family, the genetic variable $C$ is defined similarly. LIME uses the multiplicative relative risk model

$$P(D=1|M,F,C) = \delta\, R_1^{I(C=1)} R_2^{I(C=2)} \times R_{\text{im}}^{I(C=1\ \&\ \text{from mother})} S_1^{I(M=1)} S_2^{I(M=2)} \quad (1)$$

for the disease prevalence, where the parameters $R_1$ and $R_2$ denote the effect of one or two copies of an individual's own minor allele, $R_{\text{im}}$ denotes imprinting effect, $S_1$ and $S_2$ denote the effect of one or two copies of the mother's minor allele, and $\delta$ is the phenocopy rate. The indicator variable $D$ denotes the disease status of a child (1—affected; 0—normal).

To be sufficiently general to accommodate various designs, we consider nuclear families with both parents present (case or control complete families) and families for which fathers are missing (case or control incomplete families). A case (complete/incomplete) family is one for which ascertainment (the conditional event) is through an affected child, whereas a control family is ascertained through an unaffected child. Each family may contain a number of additional, nonprobands, siblings who may or may not be affected. Clearly, our ascertainment is not through a family, but through an individual (proband; single ascertainment). Therefore, our analysis will be conditional on the proband data to correct for bias (Fisher, 1934), which is different from correcting for bias for length-biased sampling. Suppose there are $N_t^1$ ($N_p^1$) and $N_t^0$ ($N_p^0$) affected and unaffected complete (incomplete) families, respectively, then $N = N_t^1 + N_t^0 + N_p^1 + N_p^0$ is the sample size, the total number of independent nuclear families.

Based on the ascertainment criterion, the proband (be it affected or unaffected) will be treated differently from those who are recruited after the family is ascertained. We use $D_1 = 1(0)$ to denote the proband being affected (unaffected). We use $D_i = 1(0)$ to denote the affection status of each affected (unaffected) sibling, $i \geq 2$. For a complete family, we use $M, F, C_1$ to denote the genotype scores (genetic variables) of the mother, father, and proband, and we use $C_i, i \geq 2$, to denote the genotype scores of additional siblings, if any. Each of such variables can take the value of 0, 1, or 2 as described earlier. Probability of the observed data from a complete family will then be conditional on the affection status of the proband only (not the other siblings):

$$P(M, F, C_1, C_i, D_i, i = 2, \cdots \mid D_1) = P(M, F, C_1|D_1) \prod_{i\geq 2} [P(D_i|M, F, C_i) P(C_i|M, F)],$$

where $D_1 = 1$ for a case family and $D_1 = 0$ for a control family, and the products over $i \geq 2$ follow from Mendel's first law, which states conditional independence of children's data given parents' genotypes. Thus, the genotype scores of the probands can be thought of as obtained from a "retrospective" design whereas the data for the additional siblings are treated as from a "prospective" design. Following the discussion in Yang and Lin (2013), for the part of data that represent a retrospective design, we can extract from the full likelihood a component (partial likelihood) that can be thought of as the products of likelihoods from a stratified prospective design (binomial kernels). This will then be combined with the prospective part of the data. Specifically, each proband and the parents form a proband–parent triad (either case–parent or control–parent triad), whereas each additional sibling (nonproband) and the parents form a sibling–parent triad (either case–sibling–parent or control–sibling–parent triad). Each triad can be classified according to their genotype configuration $(M, F, C)$. Let $n_{\text{mfc}}$ be the number of proband–parent triads with $M = m, F = f$, and $C = c$, and among such triads, $n_{\text{mfc}}^1$ and $n_{\text{mfc}}^0$ are the numbers of case–parent and control–parent triads, respectively ($n_{\text{mfc}} = n_{\text{mfc}}^1 + n_{\text{mfc}}^0$). We define $sn_{\text{mfc}}$, $sn_{\text{mfc}}^1$, and $sn_{\text{mfc}}^0$ ($sn_{\text{mfc}} = sn_{\text{mfc}}^1 + sn_{\text{mfc}}^0$) similarly for sibling–parent triads. Further, denote the vector of parameters of interest by $\boldsymbol{\theta} = (\delta, R_1, R_2, R_{\text{im}}, S_1, S_2)^\top$, and the vector of nuisance parameters (including mating type probabilities) by $\boldsymbol{\phi}$. With the fixed total of $N_t^1$ case complete families and $N_t^0$ control complete families, the likelihood from

the observed data can be written, up to a proportionality, as

$$\begin{aligned}
&\prod_{(m,f,c)} P(m,f,c|D=1)^{n^1_{\mathrm{mfc}}} P(m,f,c|D=0)^{n^0_{\mathrm{mfc}}} \\
&\quad\times P(D=1|m,f,c)^{sn^1_{\mathrm{mfc}}} P(D=0|m,f,c)^{sn^0_{\mathrm{mfc}}} \\
&\propto \left\{\prod_{(m,f,c)} (p_{\mathrm{mfc}})^{n^1_{\mathrm{mfc}}}(1-p_{\mathrm{mfc}})^{n^0_{\mathrm{mfc}}}\right\} \\
&\quad\times \left\{\prod_{(m,f,c)} (P(D=1|m,f,c))^{sn^1_{\mathrm{mfc}}}(P(D=0|m,f,c))^{sn^0_{\mathrm{mfc}}}\right\} \\
&\quad\times \left\{\prod_{(m,f,c)} [s_{\mathrm{mfc}} P(M=m,F=f,C=c)]^{n^1_{\mathrm{mfc}}+n^0_{\mathrm{mfc}}}\right\}, \qquad (2)
\end{aligned}$$

where

$$\begin{aligned}
s_{\mathrm{mfc}} \equiv s_{\mathrm{mfc}}(\boldsymbol{\theta}) &= \frac{N^1_t P(D=1|M=m,F=f,C=c)}{P(D=1)} \\
&\quad + \frac{N^0_t P(D=0|M=m,F=f,C=c)}{P(D=0)}, \\
p_{\mathrm{mfc}} \equiv p_{\mathrm{mfc}}(\boldsymbol{\theta}) &= \frac{N^1_t P(D=1|M=m,F=f,C=c)}{P(D=1)} \Big/ s_{\mathrm{mfc}}(\boldsymbol{\theta}).
\end{aligned} \qquad (3)$$

We note that the last term in (2) (i.e., term in the last set of curly brackets) is the consequence of the reparameterization, given in (3), applied to the likelihood formula given in the first line. Further, $P(D=1)$ is the disease prevalence, which can typically be retrieved from the Incidence and Prevalence Database (IPD) (`http://www.tdrdata.com/IPD/ipd_init.aspx`) or other sources.

We note that only $P(M=m,F=f,C=c)$ contains the nuisance parameters in $\boldsymbol{\phi}$. That is, the factors within the first two sets of curly brackets in (2) contain only parameters in $\boldsymbol{\theta}$ because only penetrance probabilities as defined in (1) are involved, and therefore, it is treated as the partial likelihood (Yang and Lin, 2013; Han et al., 2013). In fact, the first factor can be regarded as the likelihood representing the reorganized data conditional on each possible triad $(m,f,c)$ type. Within each type, counts of the case–parent triads and control–parent triads follow a "renormalized" binomial distribution with the following probability of being a case–parent triad:

$$\begin{aligned}
p_{\mathrm{mfc}} \equiv p_{\mathrm{mfc}}(\boldsymbol{\theta}) &= \frac{E(n^1_{\mathrm{mfc}})}{E(n^1_{\mathrm{mfc}}+n^0_{\mathrm{mfc}})} \\
&= \frac{N^1_t P(m,f,c|D=1)}{N^1_t P(m,f,c|D=1)+N^0_t P(m,f,c|D=0)} \\
&= \frac{N^1_t P(D=1|m,f,c)/P(D=1)}{N^1_t P(D=1|m,f,c)/P(D=1)+N^0_t P(D=0|m,f,c)/P(D=0)},
\end{aligned}$$

where $E(n^1_{\mathrm{mfc}})$ and $E(n^0_{\mathrm{mfc}})$ denote the expectations of observing the $(m,f,c)$ genotype configuration among the case–parent triads and control–parent triads, respectively. This manipulation turns data from a retrospective design into a "prospective" likelihood stratified according to each type. Thus, the "binomial kernel" probabilities in the first factor represent the contributions from the probands. The second factor, on the other hand, represents the contributions from the additional siblings, whose affection statuses are obtained prospectively and therefore the binomial probability is simply the penetrance probability.

Similar argument as above can be applied to incomplete families (with the exclusion of the case in which $M=1$ and $C=1$ due to ambiguity of parental genotype contribution (Yang and Lin, 2013)), leading to the following *partial log-likelihood* based on all data:

$$\begin{aligned}
l_{\mathrm{par}}(\boldsymbol{\theta}) &= \sum_{m,f,c}\left\{n^1_{\mathrm{mfc}}\times\log[p_{\mathrm{mfc}}(\boldsymbol{\theta})]+n^0_{\mathrm{mfc}}\times\log[1-p_{\mathrm{mfc}}(\boldsymbol{\theta})]\right\} \\
&\quad+\sum_{(m,c)\neq(1,1)}\left\{n^1_{\mathrm{mc}}\times\log[p_{\mathrm{mc}}(\boldsymbol{\theta})]+n^0_{\mathrm{mc}}\times\log[1-p_{\mathrm{mc}}(\boldsymbol{\theta})]\right\} \\
&\quad+\sum_{m,f,c}\left\{sn^1_{\mathrm{mfc}}\times\log[q_{\mathrm{mfc}}(\boldsymbol{\theta})]+sn^0_{\mathrm{mfc}}\times\log[1-q_{\mathrm{mfc}}(\boldsymbol{\theta})]\right\} \\
&\quad+\sum_{(m,c)\neq(1,1)}\left\{sn^1_{\mathrm{mc}}\times\log[q_{\mathrm{mc}}(\boldsymbol{\theta})]+sn^0_{\mathrm{mc}}\times\log[1-q_{\mathrm{mc}}(\boldsymbol{\theta})]\right\} \\
&= l_{t1}(\boldsymbol{\theta})+l_{p1}(\boldsymbol{\theta})+l_{t2}(\boldsymbol{\theta})+l_{p2}(\boldsymbol{\theta}),
\end{aligned}$$

where $n^1_{\mathrm{mc}}, n^0_{\mathrm{mc}}, sn^1_{\mathrm{mc}}$, and $sn^0_{\mathrm{mc}}$ are genotype counts for mother–child pairs defined similarly as for triads. Furthermore, $p_{\mathrm{mfc}}(\boldsymbol{\theta})$ and $s_{\mathrm{mfc}}(\boldsymbol{\theta})$ are as defined in (3), and

$$\begin{aligned}
s_{\mathrm{mc}}(\boldsymbol{\theta}) &= \frac{N^1_p P(D=1|M=m,C=c)}{P(D=1)} \\
&\quad+\frac{N^0_p P(D=0|M=m,C=c)}{P(D=0)}, \\
p_{\mathrm{mc}}(\boldsymbol{\theta}) &= \frac{N^1_p P(D=1|M=m,C=c)}{P(D=1)}\Big/ s_{\mathrm{mc}}(\boldsymbol{\theta}), \\
q_{\mathrm{mfc}}(\boldsymbol{\theta}) &= P(D=1|M=m,F=f,C=c), \\
q_{\mathrm{mc}}(\boldsymbol{\theta}) &= P(D=1|M=m,C=c).
\end{aligned}$$

The *effective* total sample size, $n$, in the partial log-likelihood $l_{\mathrm{par}}(\boldsymbol{\theta})$, is computed as

$$\begin{aligned}
n &= \sum_{m,f,c}[n^0_{\mathrm{mfc}}+n^1_{\mathrm{mfc}}]+\sum_{(m,c)\neq(1,1)}[n^0_{\mathrm{mc}}+n^1_{\mathrm{mc}}] \\
&\quad+\sum_{m,f,c}[sn^0_{\mathrm{mfc}}+sn^1_{\mathrm{mfc}}]+\sum_{(m,c)\neq(1,1)}[sn^0_{\mathrm{mc}}+sn^1_{\mathrm{mc}}] \\
&= (N^0_t+N^1_t+eN^0_p+eN^1_p)+(sN^0_t+sN^1_t+esN^0_p+esN^1_p) \\
&= (n_t+en_p)+(sn_t+esn_p)
\end{aligned}$$

where $(sN^0_t, sN^1_t)$ are defined similar as $(N^0_t, N^1_t)$, and are the total number of unaffected and affected siblings in all

complete families, respectively, and $eN_p^j = \sum_{(m,c)\neq(1,1)} n_{\text{mc}}^j$, and $esN_p^j = \sum_{(m,c)\neq(1,1)} sn_{\text{mc}}^j$, for $j = 0, 1$. Hence, $(n_t + en_p)$ is the total number of independent nuclear families excluding proband–mother pairs falling into the $(m, c) = (1, 1)$ category, and $(sn_t + esn_p)$ is the total number of additional siblings excluding those in incomplete families whose genotypes with the mothers falling into the $(m, c) = (1, 1)$ category.

We use the partial log-likelihood function $l_{\text{par}}(\boldsymbol{\theta})$ for statistical inference about $\boldsymbol{\theta}$. The *maximum partial likelihood estimator* (MPLE) of $\boldsymbol{\theta}$ is denoted by

$$\widehat{\boldsymbol{\theta}}_n = \text{argmax}_{\boldsymbol{\theta}} l_{\text{par}}(\boldsymbol{\theta}).$$

We assume that the MPLE is obtained by solving the score-type equation

$$\frac{\partial l_{\text{par}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = l'_{\text{par}}(\boldsymbol{\theta}) = l'_{t1}(\boldsymbol{\theta}) + l'_{p1}(\boldsymbol{\theta}) + l'_{t2}(\boldsymbol{\theta}) + l'_{p2}(\boldsymbol{\theta}) = \mathbf{0}. \quad (4)$$

### 2.2. *Asymptotic Properties*

We first introduce some additional notations. In the multiplicative relative risk model (1) for the disease prevalence, let $\boldsymbol{\theta}_0$ be the true value of the parameter $\boldsymbol{\theta} = (\delta, R_1, R_2, R_{\text{im}}, S_1, S_2)^\top$. We assume that $\boldsymbol{\theta}_0$ is an interior point of the parameter space $\boldsymbol{\Theta} \subset \mathbb{R}^6$.

As in standard likelihood theory, some regularity conditions are needed in order to study the large sample behavior of the MPLE $\widehat{\boldsymbol{\theta}}_n$. To focus on the main results, these conditions are listed in Supplementary Material A.1.1. Theorem 1 gives the large sample behavior of $\widehat{\boldsymbol{\theta}}_n$.

THEOREM 1. *Under regularity conditions R1–R5 in Supplementary Material A.1.1, we have the following:*

*(i) The score-type equation (4) has a solution $\widehat{\boldsymbol{\theta}}_n$ that is a consistent estimator of $\boldsymbol{\theta}_0$, i.e., $\widehat{\boldsymbol{\theta}}_n \longrightarrow^p \boldsymbol{\theta}_0$, as $n \to \infty$. Furthermore, the consistent solution $\widehat{\boldsymbol{\theta}}_n$ is unique.*

*(ii) Asymptotic normality: $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \longrightarrow^d N(0, \boldsymbol{I}^{-1}(\boldsymbol{\theta}_0))$, as $n \to \infty$, where the information matrix $\boldsymbol{I}(\boldsymbol{\theta}_0)$ is given by*

$$\begin{aligned}\boldsymbol{I}(\boldsymbol{\theta}_0) &= \sum_{m,f,c} \frac{[p'_{\text{mfc}}(\boldsymbol{\theta}_0)][p'_{\text{mfc}}(\boldsymbol{\theta}_0)]^\top \times B_{\text{mfc}}}{p_{\text{mfc}}(\boldsymbol{\theta}_0)(1 - p_{\text{mfc}}(\boldsymbol{\theta}_0))} \\ &\quad + \sum_{(m,c)\neq(1,1)} \frac{[p'_{\text{mc}}(\boldsymbol{\theta}_0)][p'_{\text{mc}}(\boldsymbol{\theta}_0)]^\top \times B_{\text{mc}}}{p_{\text{mc}}(\boldsymbol{\theta}_0)(1 - p_{\text{mc}}(\boldsymbol{\theta}_0)} \\ &\quad + \sum_{m,f,c} \frac{[q'_{\text{mfc}}(\boldsymbol{\theta}_0)][q'_{\text{mfc}}(\boldsymbol{\theta}_0)]^\top \times C_{mfc}}{q_{\text{mfc}}(\boldsymbol{\theta}_0)(1 - q_{\text{mfc}}(\boldsymbol{\theta}_0))} \\ &\quad + \sum_{(m,c)\neq(1,1)} \frac{[q'_{\text{mc}}(\boldsymbol{\theta}_0)][q'_{\text{mc}}(\boldsymbol{\theta}_0)]^\top \times C_{\text{mc}}}{q_{\text{mc}}(\boldsymbol{\theta}_0)(1 - q_{\text{mc}}(\boldsymbol{\theta}_0))} \\ &= \boldsymbol{I}_{t1}(\boldsymbol{\theta}_0) + \boldsymbol{I}_{p1}(\boldsymbol{\theta}_0) + \boldsymbol{I}_{t2}(\boldsymbol{\theta}_0) + \boldsymbol{I}_{p2}(\boldsymbol{\theta}_0),\end{aligned}$$

*where $p'_{\text{mc}}(\boldsymbol{\theta}_0), p'_{\text{mc}}(\boldsymbol{\theta}_0), q'_{\text{mfc}}(\boldsymbol{\theta}_0)$, and $q'_{\text{mfc}}(\boldsymbol{\theta}_0)$ are the gradients of the corresponding probabilities evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, and $0 \le B_{\text{mfc}} < 1, 0 \le B_{\text{mc}} < 1, 0 \le C_{\text{mfc}} < 1, 0 \le C_{\text{mc}} < 1$, are the limits in probability of $\frac{n_{\text{mfc}}}{n}$, $\frac{n_{\text{mc}}}{n}, \frac{sn_{\text{mfc}}}{n}$, $\frac{sn_{\text{mc}}}{n}$, respectively, when $n \to \infty$.*

The proof is given in the Supplementary Material A.1.2. Theorem 1 accommodates general cases. All the combinations of proband triads/pairs with an arbitrary number of additional siblings are covered. For part (ii), the terms $B_{\text{mfc}}$, $B_{\text{mc}}$, $C_{\text{mfc}}$, and $C_{\text{mc}}$ are zero only for the cases without proband triads, proband pairs, additional triads, or additional pairs, respectively. The calculation of these constants are provided in the Supplementary Material A.1.3.

### 2.3. *Calculation of per Family and per Individual Information Content*

The Fisher information matrix $\boldsymbol{I}(\boldsymbol{\theta}_0)$ given in Theorem 1 provides the expected information *per effective family*. To compare different study designs, we need the expected *information per family* and *per individual*, with the corresponding matrices denoted as $\boldsymbol{I}_{\text{fam}}(\boldsymbol{\theta}_0)$ and $\boldsymbol{I}_{\text{ind}}(\boldsymbol{\theta}_0)$, respectively. The calculation of these two matrices in terms of $\boldsymbol{I}(\boldsymbol{\theta}_0)$ is described as follows.

We consider the general setting of mix families each with $k$ additional siblings, where $k = 0, 1, 2, \cdots$. Let $h$ be the ratio of effective sample size $n$ to the total count of family $N$, that is $h = n/N$ (with more details given in Supplementary Material A.1.3). The expected information per family is then given by

$$\boldsymbol{I}_{\text{fam}}(\boldsymbol{\theta}_0) = \frac{n}{N} \times \boldsymbol{I}(\boldsymbol{\theta}_0) = h \times \boldsymbol{I}(\boldsymbol{\theta}_0). \quad (5)$$

Several examples are provided in the Supplementary Material A.2.1.

To compute expected information per individual, let $n^*$ be the total number of individuals involved, including all the fathers, mothers, and offsprings. Denote $g$ as the ratio of the total number of individual involved to the total number of families, that is, $g = n^*/N$. Then, the expected information per individual is

$$\boldsymbol{I}_{\text{ind}}(\boldsymbol{\theta}_0) = \frac{n}{n^*} \times \boldsymbol{I}(\boldsymbol{\theta}_0) = \frac{h}{g} \times \boldsymbol{I}(\boldsymbol{\theta}_0) = \frac{\boldsymbol{I}_{\text{fam}}(\boldsymbol{\theta}_0)}{g}. \quad (6)$$

An alternative representation (interpretation) and more examples are given in Supplementary Material A.2.2.

### 2.4. *Numerical Study: Empirical versus Asymptotic Variances*

We first use extensive simulations under a variety of disease models, scenarios, and small to large sample sizes to verify the asymptotic properties of the procedure empirically. We see from Table S1 and Figures S1–S8 in the Supplementary Material, as the sample size increases, the relative differences between a parameter and its corresponding MPLE get closer and closer to zero. Further, the empirical distributions of the relative differences are not distinguishable from a normal distribution based on a statistical test, as the sample size gets larger. The full details are given in Supplementary Material A.3.

To evaluate how well the asymptotic variances (diagonal elements of $\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0)$) can approximate actual variances in finite

**Table 1**
*Eight disease models represented by relative risks and eight scenarios comprised of three factors*

A. Disease models

| | Relative risk | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Para.[a] | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| $R_1$ | 1 | 2 | 1 | 1 | 1 | 3 | 1 | 3 |
| $R_2$ | 1 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| $R_{\text{im}}$ | 1 | 1 | 1 | 1 | 3 | 1/3 | 3 | 1/3 |
| $S_1$ | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 2 |
| $S_2$ | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 2 |

B. Scenarios

| | Factor value | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Factor[b] | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| MAF | 0.1 | 0.1 | 0.1 | 0.1 | 0.3 | 0.3 | 0.3 | 0.3 |
| PREV | 0.05 | 0.05 | 0.15 | 0.15 | 0.05 | 0.05 | 0.15 | 0.15 |
| HWE | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |

[a] $R_1$: relative risk of carrying one variant allele;
$R_2$: relative risk of carrying two variant alleles;
$R_{\text{im}}$: imprinting effect parameter with a single variant allele from mother;
$S_1$: maternal effect with mother carrying one variant allele;
$S_2$: maternal effect with mother carrying two variant alleles.

[b] MAF: minor allele frequency;
PREV: prevalence (rare = 0.05; common = 0.15);
HWE: Hardy–Weinberg equilibrium (Yes = 1; No = 0).

Note that a specification of a disease model and a scenario completely determines the phenocopy rate $\delta$ and thus the penetrance model in (1).

samples, an important issue for considering study designs with finite sample sizes, we compare the two in a variety of combinations of disease models, scenarios, and sample sizes. Specifically, we consider eight disease models as given in Table 1A. Note that the first model is a null setting with no genetic effect ($R_1 = R_2 = R_{\text{im}} = S_1 = S_2 = 1$). Under each model, we investigate eight combinations (scenarios; Table 1B) of three factors: minor allele frequency (MAF) $\{0.1, 0.3\}$, population disease prevalence $P(D = 1)$ (PREV) $\{0.05, 0.15\}$, and whether Hardy–Weinberg equilibrium (HWE) holds (no = 0, yes = 1). Suppose $p$ is the MAF, then when HWE holds, the probabilities of a genotype score being 0, 1, and 2 are $(1-p)^2$, $2(1-p)p$, and $p^2$, respectively. When HWE does not hold, the probabilities are $(1-p)^2(1-\zeta) + (1-p)\zeta$, $2p(1-p)(1-\zeta)$, and $p^2(1-\zeta) + p\zeta$, where $\zeta$ is the inbreeding parameter (Weir, 1996), which in our simulation is set to be 0.1 and 0.3 for males and females, respectively. With the specification of each scenario and a disease model, the phenocopy rate $\delta$, and consequently the penetrance probability (1) are fully specified. Note that these eight combinations of scenarios are chosen to compare and contrast the asymptotic behavior of LIME in easier situations (larger MAF/common disease/HWE) with harder ones (smaller MAF/rare disease/HWE does not hold).

We examine a total of nine data types: $\{P, M, T, P+1, M+1, T+1, P+2, M+2, T+2\}$, where "$P$" refers to the setting in which all families in the sample are of "pair type" with the father's genotype missing; "$T$" refers to the setting in which all families in the sample are of "triad type" with both parents' genotype present, and "$M$" is a mixture of "$T$" and "$P$" with the missing rate for father being 0.5 and 0.7 in affected and unaffected families, respectively, in our simulation. The number after each letter designation (if any) is the number of additional siblings (in addition to the proband) in each nuclear family. For instance, data type $T+2$ refers to a sample of families each with two parents, an affected/unaffected proband, and two additional siblings who may or may not be affected. In other words, each family is a complete nuclear family with three children. A family is labeled as a case/control family in our sample if the first child simulated is affected/unaffected with the disease (the proband) regardless of the affection status of the subsequent siblings. This is to mimic the single ascertainment scheme in real genetic studies. Note that the "first child" simulated does not necessarily have to be the "first born"; rather, it is the first child that has come to the attention (of a physician), and through whom the family is recruited for the study. We repeat the process of simulating each family until the desired numbers of families of both types are met. The sample size $N$ is set to be 200, 1000, 2000, and 10,000, with an equal number of case and control families. The results are based on 500 replications, and the variance of the estimates across the replications gives the empirical variance.

Figure 1 provides plots of differences between empirical and asymptotic variances of parameter estimators for four data types: $P$, $P+2$, $T$, and $T+2$, presented in four blocks. Within each block, we show results for two sample sizes and four scenarios. In each plot, there are 40 points corresponding to the

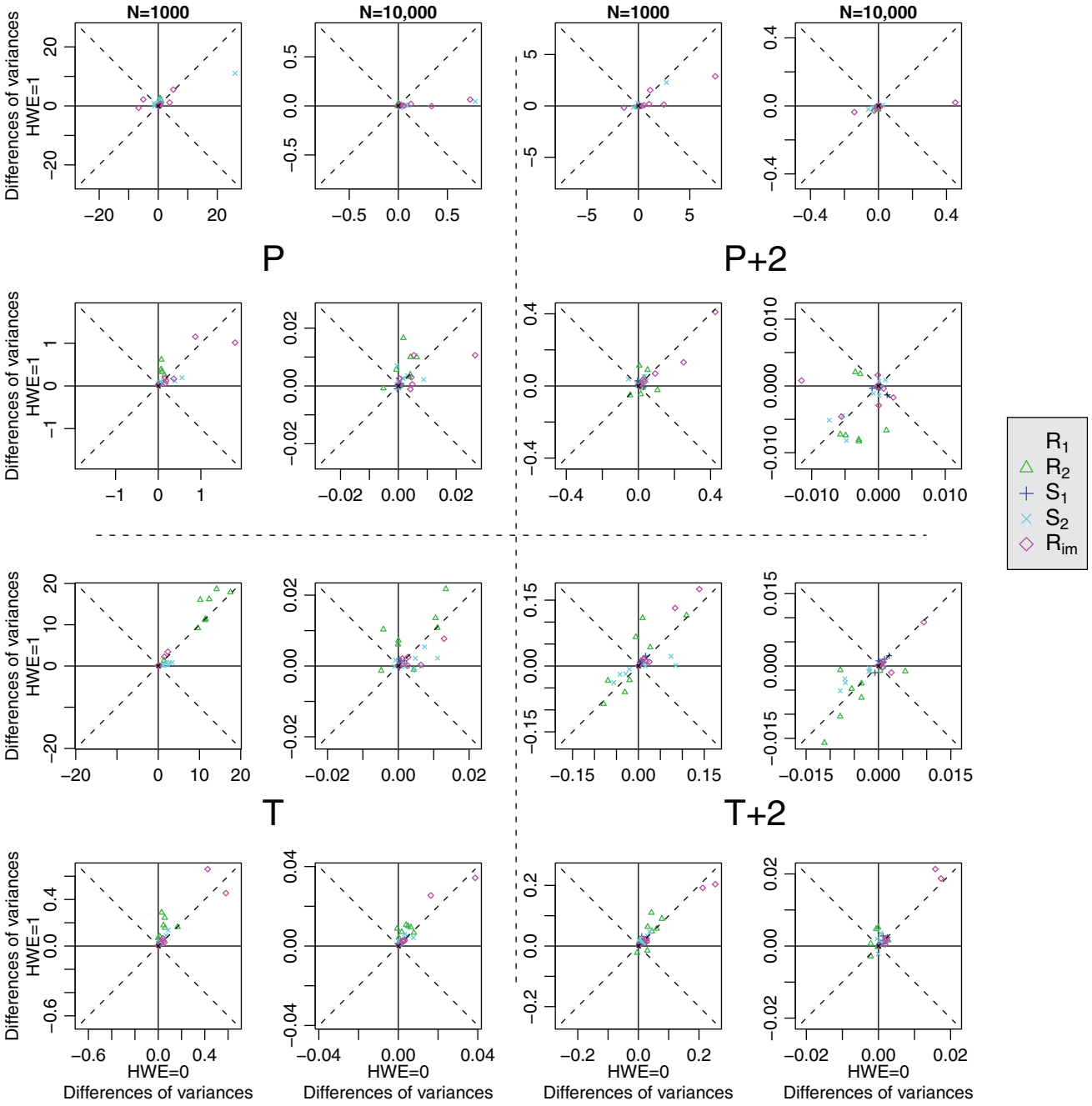


**Figure 1.** The difference between empirical and asymptotic variances for HWE = 1 versus HWE = 0 for four data types (four blocks, each with four sub-figures): top left—$P$; top right—$P + 2$; bottom left—$T$; bottom right—$T + 2$. For each data type (within each block), we show results for two sample sizes (N=1000 and N=10,000) and two scenarios: top row—MAF= 0.1 and PREV= 0.15; bottom row—MAF= 0.3, and PREV= 0.05. This figure appears in color in the electronic version of this article.

five parameters under the eight disease models. Results for all combinations of sample sizes, scenarios, and data types investigated are given in Supplementary Figures S9–S17 and are summarized in Tables 2 and 3. From the figures, one can see that, as the sample size increases, all differences get closer and closer to zero. Most of the points fall around the diagonal line, showing that the difference between whether the HWE assumption holds or not is minor, substantiating the property that LIME is robust to departure from the HWE assumption. Further, as we can see from Table 2, regardless of whether HWE hold or not, $R_2$ and $S_2$ are much harder to estimate precisely compared to $R_1$ and $S_1$ for a fixed sample size, especially for the smaller MAF scenarios, as there are fewer people who are homozygous for the minor allele. Likewise, $R_{\text{im}}$ is also more difficult to estimate due to a smaller count of mother–child pairs that are informative for estimating the

**Table 2**

*Average differences[a] between empirical and asymptotic variances of the parameter estimators for each of the eight scenarios and four sample sizes.*

| | | MAF= 0.1&PREV= 0.05 | | | | MAF= 0.1&PREV= 0.15 | | | | MAF= 0.3&PREV= 0.05 | | | | MAF= 0.3&PREV= 0.15 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 200 | 1000 | 2000 | 10,000 | 200 | 1000 | 2000 | 10,000 | 200 | 1000 | 2000 | 10,000 | 200 | 1000 | 2000 | 10,000 |
| HWE=0 | $R_1$ | 0.81 | 0.11 | 0.05 | 0.01 | 0.35 | 0.07 | 0.04 | 0.01 | 0.58 | 0.09 | 0.04 | 0.01 | 0.38 | 0.07 | 0.03 | 0.01 |
| | $R_2$ | 73.15 | 5.10 | 1.39 | 0.20 | 12.32 | 2.59 | 0.92 | 0.09 | 8.46 | 0.65 | 0.29 | 0.06 | 3.86 | 0.43 | 0.20 | 0.03 |
| | $S_1$ | 1.16 | 0.11 | 0.05 | 0.01 | 0.43 | 0.08 | 0.04 | 0.01 | 0.63 | 0.09 | 0.04 | 0.01 | 0.37 | 0.06 | 0.03 | 0.01 |
| | $S_2$ | 39.85 | 4.39 | 1.14 | 0.07 | 16.21 | 0.74 | 0.26 | 0.03 | 3.17 | 0.24 | 0.11 | 0.02 | 1.24 | 0.14 | 0.07 | 0.01 |
| | $R_{\text{im}}$ | 49.01 | 5.06 | 1.30 | 0.11 | 9.39 | 0.93 | 0.40 | 0.05 | 9.30 | 0.47 | 0.20 | 0.04 | 4.25 | 0.29 | 0.13 | 0.02 |
| HWE=1 | $R_1$ | 0.91 | 0.11 | 0.06 | 0.01 | 0.36 | 0.08 | 0.04 | 0.01 | 0.68 | 0.10 | 0.05 | 0.01 | 0.42 | 0.07 | 0.03 | 0.01 |
| | $R_2$ | 58.05 | 3.43 | 0.97 | 0.15 | 8.27 | 2.07 | 0.76 | 0.07 | 4.14 | 0.43 | 0.21 | 0.04 | 2.18 | 0.31 | 0.14 | 0.02 |
| | $S_1$ | 0.61 | 0.08 | 0.04 | 0.01 | 0.31 | 0.06 | 0.03 | 0.00 | 0.52 | 0.07 | 0.04 | 0.01 | 0.32 | 0.06 | 0.03 | 0.01 |
| | $S_2$ | 47.42 | 9.68 | 4.93 | 0.25 | 8.94 | 1.37 | 0.76 | 0.08 | 5.28 | 0.25 | 0.12 | 0.02 | 1.57 | 0.15 | 0.07 | 0.01 |
| | $R_{\text{im}}$ | 37.98 | 13.49 | 7.48 | 0.31 | 7.84 | 1.42 | 0.86 | 0.13 | 11.80 | 0.48 | 0.22 | 0.04 | 4.14 | 0.32 | 0.13 | 0.02 |

[a]Each number in the table is averaged over eight models and nine data types.

parameters. It is also apparent from the table that it is easier to estimate model parameters for diseases with a larger prevalence, as the differences are smaller compared to those with a smaller prevalence. Comparing across all nine different data types (Table 3), one can see that, as expected, the $T$ data types, consisting of all complete families and thus more information, have smaller differences between the empirical and asymptotic variances for a fixed sample size. Furthermore, additional siblings provide extra information.

## 3. Study Design Consideration

Results from the above numerical studies and those presented in Supplementary Material A.3 show that, regardless of the data type, parameter estimates will be close to the true parameter values for a large enough sample size. However, in any real study setting, resources are finite, therefore, it is important that one chooses a study design that is efficient and practicable. To address this issue, we compare nine study designs (the nine data types in our numerical study) through consideration of information content per family and per individual (in the next two subsections). We limit ourselves to only nine data types for easy presentation but the conclusion is more generally applicable to adding any number of siblings as we discuss below. We also perform sample size calculation, with some general observations summarized here and detailed results presented in Supplementary Material A.4 and Supplementary Tables S2–S9. The sample size needed to achieve a certain precision is the smallest for the T + 2 study design, whereas the P design requires the largest sample size, as one would expect. It is also seen that the homozygous genetic effect ($R_2$), homozygous maternal effect ($S_2$), and the imprinting parameter ($R_{\text{im}}$) are typically more difficult to estimate accurately, as there are fewer families informative for these parameters, for example, mother being homozygous for the minor allele.

### 3.1. *Information Content per Family*

The information content per family is computed based on (5); see Supplementary Material A.2 for the formulas for calculating such quantities for different data types. It is clear from the simulation study that it is advantageous to have complete families and additional siblings. To more clearly delineate this

**Table 3**

*Average differences[a] between empirical and asymptotic variances of the parameter estimators for each of the nine data types, four combinations of scenarios[b], and four sample sizes.*

| | MAF= 0.1&PREV= 0.05 | | | | MAF= 0.1&PREV= 0.15 | | | | MAF= 0.3&PREV= 0.05 | | | | MAF= 0.3&PREV= 0.15 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 200 | 1000 | 2000 | 10000 | 200 | 1000 | 2000 | 10000 | 200 | 1000 | 2000 | 10000 | 200 | 1000 | 2000 | 10000 |
| T | 36.17 | 1.87 | 0.49 | 0.07 | 6.83 | 3.15 | 1.01 | 0.04 | 2.99 | 0.26 | 0.12 | 0.02 | 1.88 | 0.35 | 0.16 | 0.02 |
| T+1 | 10.59 | 0.64 | 0.26 | 0.05 | 2.51 | 0.24 | 0.11 | 0.02 | 1.71 | 0.20 | 0.09 | 0.02 | 0.84 | 0.11 | 0.05 | 0.01 |
| T+2 | 5.67 | 0.42 | 0.20 | 0.04 | 1.55 | 0.16 | 0.08 | 0.01 | 1.27 | 0.16 | 0.08 | 0.01 | 0.62 | 0.08 | 0.04 | 0.01 |
| M | 48.27 | 2.18 | 0.57 | 0.08 | 8.52 | 0.65 | 0.29 | 0.05 | 4.91 | 0.33 | 0.15 | 0.03 | 3.08 | 0.24 | 0.11 | 0.02 |
| M+1 | 13.19 | 0.74 | 0.30 | 0.05 | 3.05 | 0.27 | 0.13 | 0.02 | 2.64 | 0.25 | 0.12 | 0.02 | 1.16 | 0.14 | 0.07 | 0.01 |
| M+2 | 8.11 | 0.49 | 0.23 | 0.04 | 1.99 | 0.18 | 0.09 | 0.02 | 1.82 | 0.20 | 0.10 | 0.02 | 0.80 | 0.10 | 0.05 | 0.01 |
| P | 84.28 | 16.74 | 7.52 | 0.38 | 17.25 | 2.10 | 1.11 | 0.16 | 13.87 | 0.53 | 0.23 | 0.04 | 4.98 | 0.34 | 0.15 | 0.03 |
| P+1 | 41.95 | 7.68 | 3.66 | 0.18 | 10.37 | 0.96 | 0.52 | 0.07 | 6.86 | 0.37 | 0.17 | 0.03 | 2.04 | 0.19 | 0.09 | 0.02 |
| P+2 | 29.80 | 6.66 | 2.45 | 0.13 | 5.90 | 0.74 | 0.35 | 0.05 | 4.03 | 0.29 | 0.14 | 0.02 | 1.45 | 0.14 | 0.07 | 0.01 |

[a]Each number in the table is averaged over eight models, five parameters, and two HWE levels.
[b]Each combination is by collapsing the two HWE levels with the same MAF and PREV.

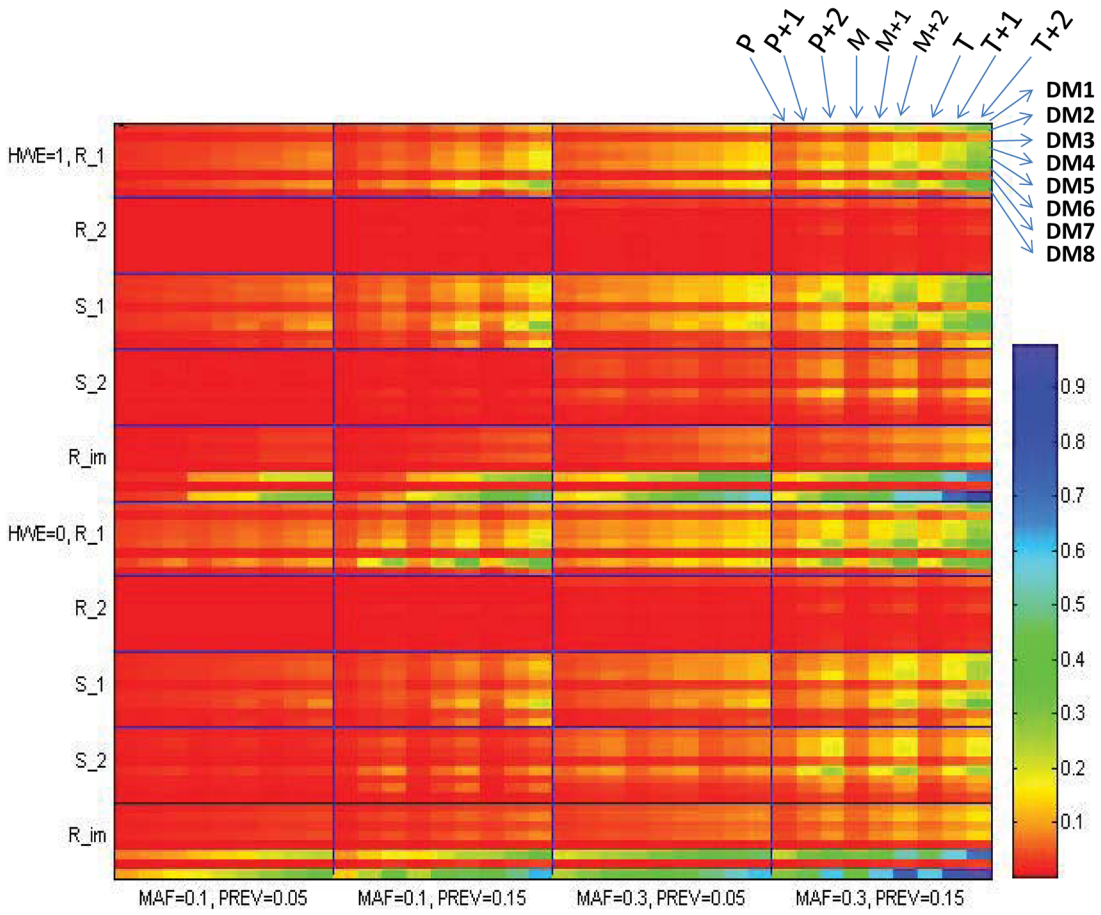


**Figure 2.** Information content per family for parameter estimation from the nine study designs (data types): $\{P, P+1, P+2, M, M+1, M+2, T, T+1, T+2\}$. Each of the four column blocks represent an MAF and PREV combination. Within each block, information from the nine data types are presented in the order as indicated in the figure. In the top half, each of the five blocks provide information for the estimation of each of the five parameters under HWE. Furthermore, each of the eight rows within a row block represent the eight disease models (DMs) in the order as indicated. The bottom half provides the same information but with the HWE assumption being violated. This figure appears in color in the electronic version of this article.

advantage from a theoretical point of view, we show, in Figure 2, the expected information content from a single family for estimating the five parameters. The eight combinations of MAF, PREV, and HWE are organized into two sets of row blocks (top and bottom) and four column blocks. Each column block contains information for nine study designs, with ordering indicated in the figure. The five subblocks within each set of the two row blocks correspond to the five parameters. Furthermore, each of the eight rows within each block are for the eight models as given in Table 1A. As expected, the amount of information increases from left to right (Figure 2) within each of the four column blocks, indicating that a complete family contains more information than an incomplete one when father's genotype is missing, and therefore, the information content for a mixed type is in-between. Additional siblings also increase the family information content. We can also see from the figure that increasing MAF from 0.1 to 0.3 and/or PREV from 0.05 to 0.15 enriches the information contained in the sample for estimating the parameters. The eight models also exhibit differences, although to a lesser extent than the study design. In general, there tends to be greater information for estimating $R_1$ and $S_1$ than for $R_2$ and $S_2$. The information for estimating the imprinting parameter, $R_{\text{im}}$, is especially model dependent, with particularly strong information for models 6 and 8, which portraits strong maternal imprinting and association effects. We note that, although there can be large discrepancy in the empirical and asymptotic variances in small samples (see first column of Figures S9–S17), especially for estimating $R_2$, $R_{\text{im}}$, and $S_2$, the use of information content remains a reasonable way of evaluating design efficiency since the patterns of discrepancy are similar across the different designs considered.

### 3.2. *Information Content per Individual*

In practice, resources are fixed, such as labor, time, equipment, and fund, which can only permit genotyping a limited number of individuals in a study. Thus, it is important to decide how to distribute the resources. To this end, we consider the information provided by a single individual for each of the nine study designs, thereby taking the size of each fam-

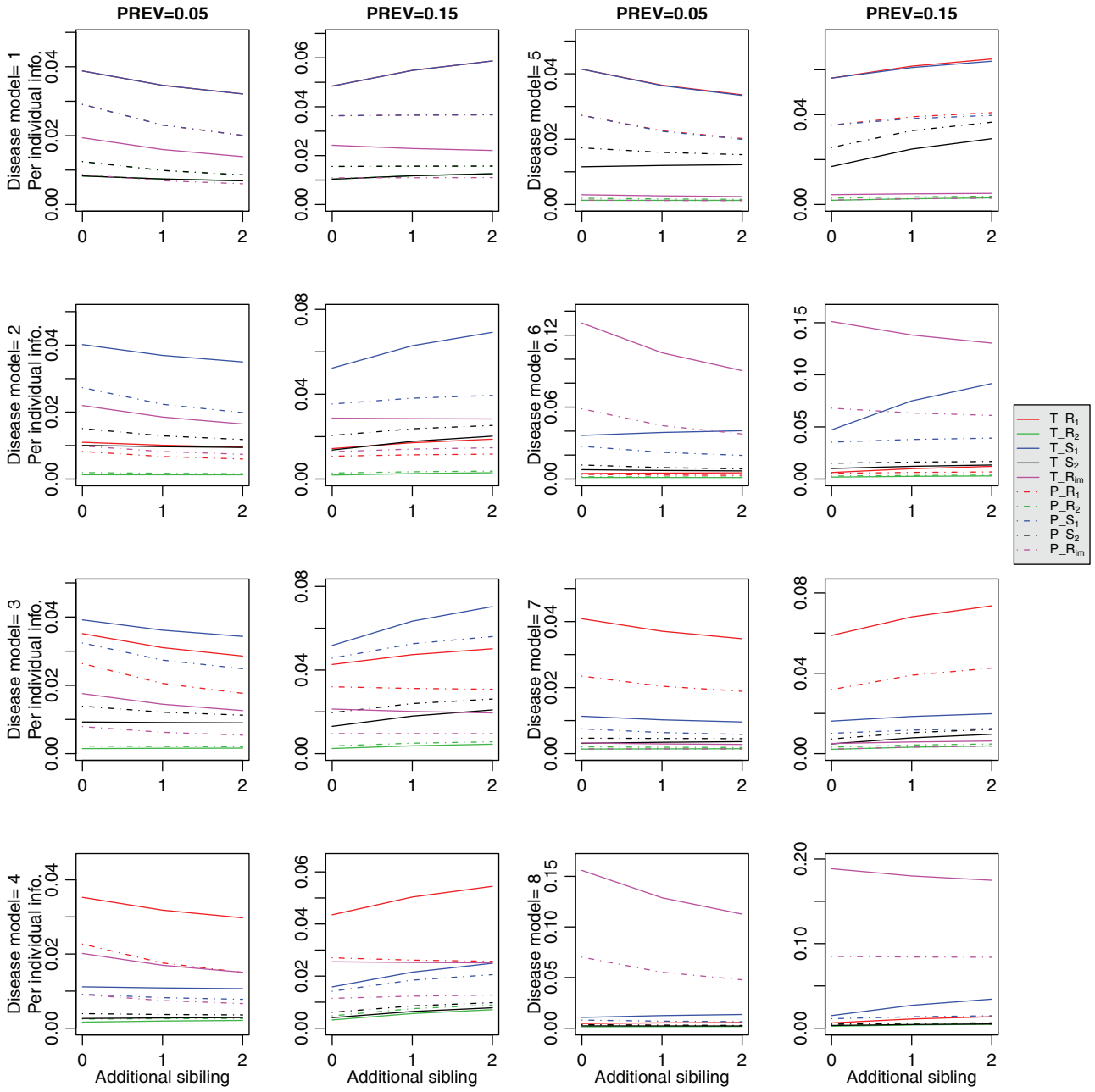


**Figure 3.** Information content per individual for eight disease models and two PREVs with MAF = 0.3 and HWE = 1. Each curve provides the information for estimating one of the five parameters, for a particular family type with 0, 1, or with 2 additional siblings. The study designs considered are the *T* and *P* data types. This figure appears in color in the electronic version of this article.

ily (genotyping cost) into account. This is asymptotic information, thus there is no need to specify a sample size. This information is particularly useful for designing a study that only has resources for genotyping a fixed number of individuals. Figure 3 shows the information content per individual for six study designs, the *P* and *T* types, when HWE holds and MAF is 0.3 (scenarios 6 and 8 in Table 1B). Plots for the other three study designs, the *M* types, are given as Supplementary Figure S18. We only show results for these six types in Figure 3 to make the contents more easy to digest without loss of generality, as the figures show that the information content per individual in an *M*-type family is always in between that of the corresponding *P* and *T* types. As can be seen from Figure 3, information content per individual is always higher in a triad family than in a pair family with the same number of siblings for estimating any of the parameters. Therefore, it is worth the extra effort to recruit both parents if at all possible.

On the other hand, one striking feature is that including additional siblings may or may not lead to a greater amount of information when genotyping cost is taken into account, re-

gardless of whether it is a complete or an incomplete family. Whether it is beneficial or not to recruit additional siblings depends on whether the additional information contributed by a sibling is greater than the average information contributed by an individual in a family with only parents (or mother) and probands. More precisely, suppose $I_T$ is the per individual information per triad family, and $I_S$ is the additional information contributed by an additional sibling, then the per individual information of a triad $+k$ sibling ($k \geq 1$) family is greater than a per individual information for a triad only family if and only if $I_S > I_T$. This is similarly true for a pair family. Therefore, if the average information for a proband and his or her parent(s) is higher than the extra information gained by adding a single sibling, the average individual information will decrease by recruiting additional siblings. Conversely, if the average information for a proband and the parent(s) is lower, we can take advantage by recruiting additional siblings. From Figure 3, we can see that, for a disease with low prevalence (PREV= 0.05), having larger families will in fact be counter productive since each additional individual does not contribute much to the estimation. On the other hand, for a relatively more common disease (PREV= 0.15), recruiting larger families is more efficient. This makes sense intuitively as both cases and controls are likely to be present in the additional siblings if a disease is common, whereas most likely only unaffected siblings will be recruited if the disease is rare. These observations are consistent with the limited simulation study presented in Han et al. (2013), in which the authors only considered PREV= 0.15 and concluded that larger families are more cost effective than families with probands only. Nevertheless, our results provide a comprehensive view of the situation, aided by the asymptotic theory. The take-home message is that which study design is suitable for a particular study depends on the (hypothesized) characteristics of the disease, with the population prevalence (which is typically available) being the most important factor, although the underlying disease model may play a role as well. Results for the other scenarios (1–5 and 7) lead to the same conclusion; all results are summarized in a heatmap (Supplementary Figure S19). To sum up, the conclusion drawn in Han et al. (2013) is only partially true. Aided by the asymptotic results, we can draw a more definitive conclusion: it is not always advantageous to recruit additional siblings; additional siblings can increase the efficiency of a study only when the disease being investigated is sufficiently common.

## 4. Discussion

In this article, we present a methodology for investigating, in a family-based design for detecting imprinting and maternal effects, whether it is better to recruit bigger families or smaller ones, by keeping the total number of individuals for genotyping to be the same. With the availability of large-scaled genotype data, case-control-family-based designs are considered to be a new paradigm for genetic epidemiology research (Hopper, 2003). Breast cancer research is one example where case-control family designs have been used (Becher et al., 2003). Studies of autism, binge eating disorder, and inflammatory bowel disease are other examples where case-control family designs have been utilized (Bolton et al., 1994; Javaras et al., 2008; Li et al., 2014). The method proposed in this article will be useful in aiding researchers in planning efficient designs to achieve desired estimation accuracy. Specifically, we demonstrate that this work offers a practical strategy for investigators to select the optimum study design within a case-control family scheme for a specific disease model before data collection. Although this work focuses on the LIME method for detecting imprinting and maternal effects, the strategy can be more generally applicable to other family-based designs, such as those based on the parent-asymmetry test (Zhou et al., 2009) or those for quantitative traits (He et al., 2011; Koning et al., 2002; Sung and Rao, 2008).

The cost consideration and some technical issues deserve further elaboration and discussion. Our conclusion on an efficient study design was based on the average (per individual) information content, which is related to genotyping cost. However, in any practical situation, there are more factors that should be considered when selecting an efficient study design. Genotyping cost is just one of the important attributes; phenotyping and family recruitment can be more expensive because of availability of cost-effective large-scale genotyping techniques. As such, if additional siblings are available, it would still be beneficial to recruit them, as LIME can be applied to a sample with a mixture of different data types.

Recall that in our partial likelihood formulation, case–mother and control–mother pairs with genotype combination $(1, 1)$ are excluded due to ambiguity of parental genotype contribution. This exclusion may lead to potential power loss (Yang and Lin, 2013), but not bias. This is because LIME turns a retrospective design into a prospective one through conditioning on each combination of genotype pairs (for proband–mother pair data). As such, data for each genotype combination (with combined data from case and control families) contribute independently to the partial likelihood. What is important is the "relative proportions" of case–mother/control–mother pairs within each genotype combination. As such, deleting proband–mother pairs with $(1, 1)$ genotypes will not lead to bias. Also, as we pointed out earlier, population prevalence for common diseases can typically be obtained from databases. Nevertheless, we evaluated the effects of misspecification of prevalence by as much as 20% over, or under, the true value. We can see, from Figure S20, that the powers and type I errors closely track those with the correct specification, demonstrating robustness of the LIME procedure with moderate departure from population prevalence.

As we saw in Figure 1 and Supplementary Figures S9–S17, parameter estimates (especially for $R_2$, $R_{\text{im}}$, and $S_2$) can be far from their true values, due to a flat partial likelihood surface. As such, initial values are important. Other than the typical recommendation of multiple initial values, a strategy that works well in our study is the use of estimates from a subset as the starting point for the full data sets to obtain accurate estimates. The idea is that a smaller data set can more easily identify the neighborhood where the maximizer resides, whereas a larger data set can provide greater amount of information to find the maximizer itself. Alternatively, one may consider a regularized partial likelihood to rein in any potentially wild estimates, although this is out of the scope of this article.

## 5. Supplementary Material

Web Appendices, Tables, and Figures, referenced in Sections 2, 3, and 4, and an R package for calculating information and sample size, may be accessed at the *Biometrics* website on Wiley Online Library.


## Acknowledgements

We thank the Editor, Associate Editor, and two anonymous reviewers for their constructive comments and suggestions. This research was partially supported by the National Science Foundation grant DMS-1208968, and the Natural Sciences and Engineering Research Council of Canada.


## References


Becher, H., Schmidt, S., and Chang-Claude, J. (2003). Reproductive factors and familial predisposition for breast cancer by age 50 years. A case-control-family study for assessing main effects and possible gene–environment interaction. *International Journal of Epidemiology* **32**, 38–48.

Bolton, P., Macdonald, H., and Pickles, A. (1994). A case-control family history study of autism. *Journal of Child Psychology and Psychiatry* **35**, 877–900.

Fisher, R. A. (1934). The effect of methods of ascertainment upon the estimation of estimation of frequencies. *Annals of Human Genetics* **6**, 13–25.

Han, M., Hu, Y. Q., and Lin, S. (2013). Joint detection of association, imprinting and maternal effects using all children and their parents. *European Journal of Human Genetics* **27**, 1449–1456.

He, F., Zhou, J. Y., Hu, Y. Q., Sun, F., Yang, J., Lin, S., and Fung, W. K. (2011). Detection of parent-of-origin effects for quantitative traits in complete and incomplete nuclear families with multiple children. *American Journal of Epidemiology* **174**, 226–233.

Hopper, J. L. (2003). Commentary: Case-control-family designs: A paradigm for future epidemiology research? *International Journal of Epidemiology* **32**, 48–50.

Javaras, K. N., Laird, N. M., Reichborn-Kjennerud, T., Bulik, C. M., Pope, H. G., and Hudson, J. I. (2008). Familiality and heritability of binge eating disorder: Results of a case-control family study and a twin study. *International Journal of Eating Disorders* **41**, 174–179.

Koning, D., Bovenhuis, H., and Arendonk, J. A. M. (2002). On the detection of imprinted quantitative trait loci in experimental crosses of outbred species. *Genetics Society of America* **161**, 931–938.

Lawson, H. A., Cheverud, J. M., and Wolf, J. B. (2013). Genomic imprinting and parent-of-origin effects on complex traits. *Nature Reviews Genetics* **14**, 609–17.

Li, G. and Cui, Y. (2010). A general statistical framework for dissecting parent-of-origin effects underlying endosperm traits in flowering plants. *The Annals of Applied Statistics* **4**, 1214–1233.

Li, X., Sui, Y., Liu, T., Wang, J., Li, Y., Lin, Z., Hegarty, J., Koltun, W., Wang, Z., and Wu, R. (2014). A model for family-based case-control studies of genetic imprinting and epistasis. *Briefings in Bioinformatics* **15**, 1069–1079.

Lindsay, B. G. (1980). Nuisance parameters, mixture models, and the efficiency of partial likelihood estimators. *Philosophical Transactions of the Royal Society of London* **A 296**, 639–662.

Nousome, D., Lupo, P. J., Okcu, M. F., and Scheurer, M. E. (2013). Maternal and offspring xenobiotic metabolism haplotypes and the risk of childhood acute lymphoblastic leukemia. *Leukemia Research* **37**, 531–5.

Sung, Y. J. and Rao, D. C. (2008). Model-based linkage analysis with imprinting for quantitative traits: Ignoring imprinting effects can severely jeopardize detection of linkage. *Genetic Epidemiology* **32**, 487–496.

Weir, B. S. (1996). *Genetic Data Analysis II* Summit: Sinauer.

Yang, J. and Lin, S. (2013). Robust partial likelihood approach for detecting imprinting and maternal effects using case-control families. *Annals of Applied Statistics* **1**, 249–268.

Zhou, J. Y., Hu, Y. Q., Lin, S., and Fung, W. K. (2009). Detection of parent-of-origin effects based on complete and incomplete nuclear families with multiple affected children. *Human Heredity* **67**, 1–12.


*Received October* 2014. *Revised June* 2015. *Accepted July* 2015.