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_{im} denotes imprinting effect, S_1 and S_2 denote the effect of one or two copies of the mother's minor allele, and δ 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_{mfc} be the number of proband–parent triads with $M = m, F = f,$ and $C = c$, and among such triads, n_{mfc}^1 and n_{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_{mfc}, sn_{mfc}^1, and sn_{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:

$$\begin{aligned}
l_{\mathrm{par}}(\boldsymbol{\theta}) &= \sum_{m,f,c}\left\{n^1_{\mathrm{mfc}}\times\log[p_{\mathrm{mfc}}(\boldsymbol{\theta})]+n^0_{\mathrm{mfc}}\times\log[1-p_{\mathrm{mfc}}(\boldsymbol{\theta})]\right\} \\
&\quad+\sum_{(m,c)\neq(1,1)}\left\{n^1_{\mathrm{mc}}\times\log[p_{\mathrm{mc}}(\boldsymbol{\theta})]+n^0_{\mathrm{mc}}\times\log[1-p_{\mathrm{mc}}(\boldsymbol{\theta})]\right\} \\
&\quad+\sum_{m,f,c}\left\{sn^1_{\mathrm{mfc}}\times\log[q_{\mathrm{mfc}}(\boldsymbol{\theta})]+sn^0_{\mathrm{mfc}}\times\log[1-q_{\mathrm{mfc}}(\boldsymbol{\theta})]\right\} \\
&\quad+\sum_{(m,c)\neq(1,1)}\left\{sn^1_{\mathrm{mc}}\times\log[q_{\mathrm{mc}}(\boldsymbol{\theta})]+sn^0_{\mathrm{mc}}\times\log[1-q_{\mathrm{mc}}(\boldsymbol{\theta})]\right\} \\
&= l_{f1}(\boldsymbol{\theta})+l_{p1}(\boldsymbol{\theta})+l_{f2}(\boldsymbol{\theta})+l_{p2}(\boldsymbol{\theta}),
\end{aligned}$$

where $n^1_{\mathrm{mc}}, n^0_{\mathrm{mc}}, sn^1_{\mathrm{mc}}$, and sn^0_{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_{mfc}, B_{mc}, C_{mfc}, and C_{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_{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_{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 δ 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 ζ 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 δ, 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

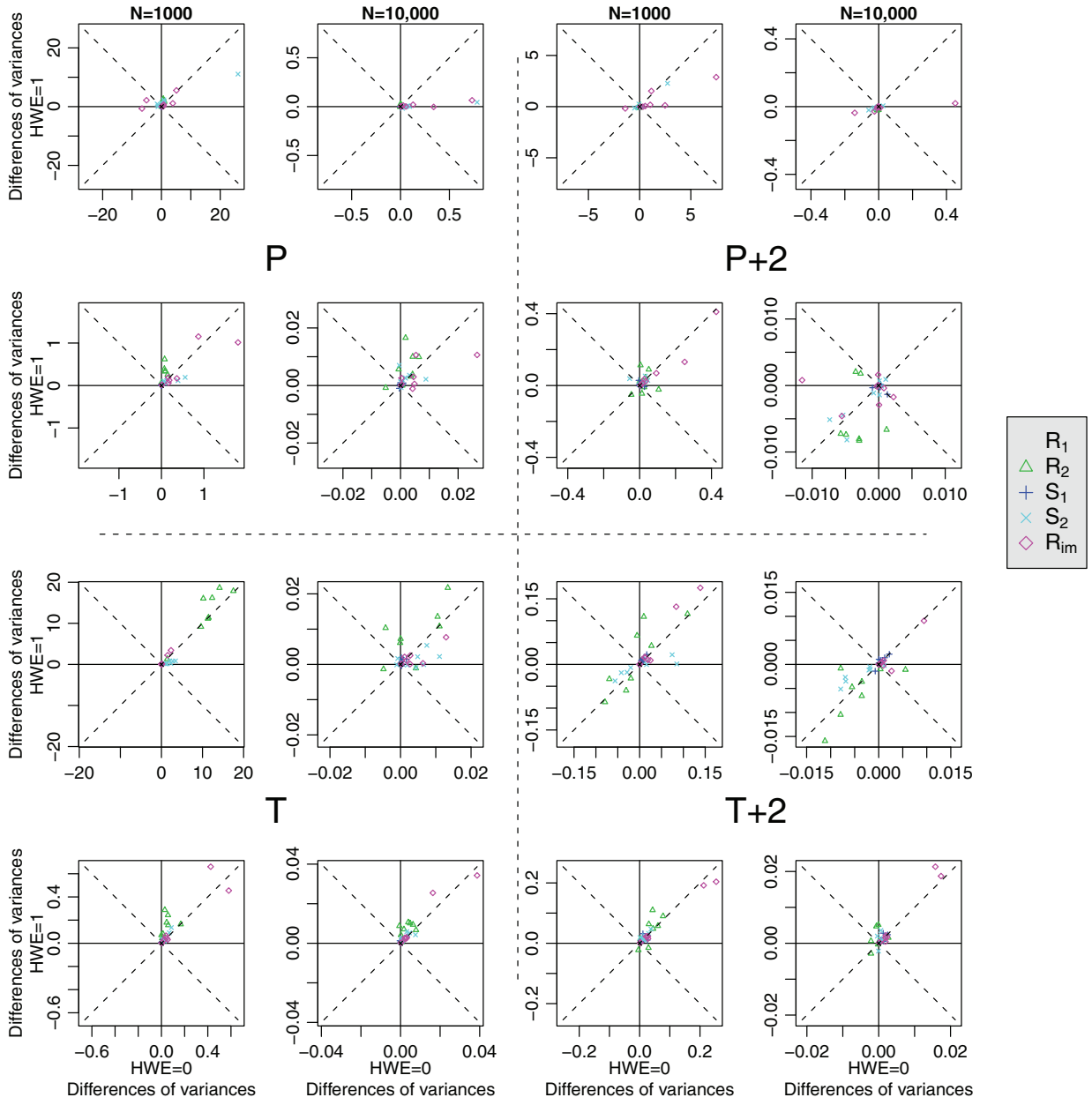


Figure 1. The difference between empirical and asymptotic variances for HWE = 1 versus HWE = 0 for four data types (four blocks, each with four sub-figures): top left—P; top right—$P + 2$; bottom left—T; bottom right—$T + 2$. For each data type (within each block), we show results for two sample sizes (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_{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_{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_{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_{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.

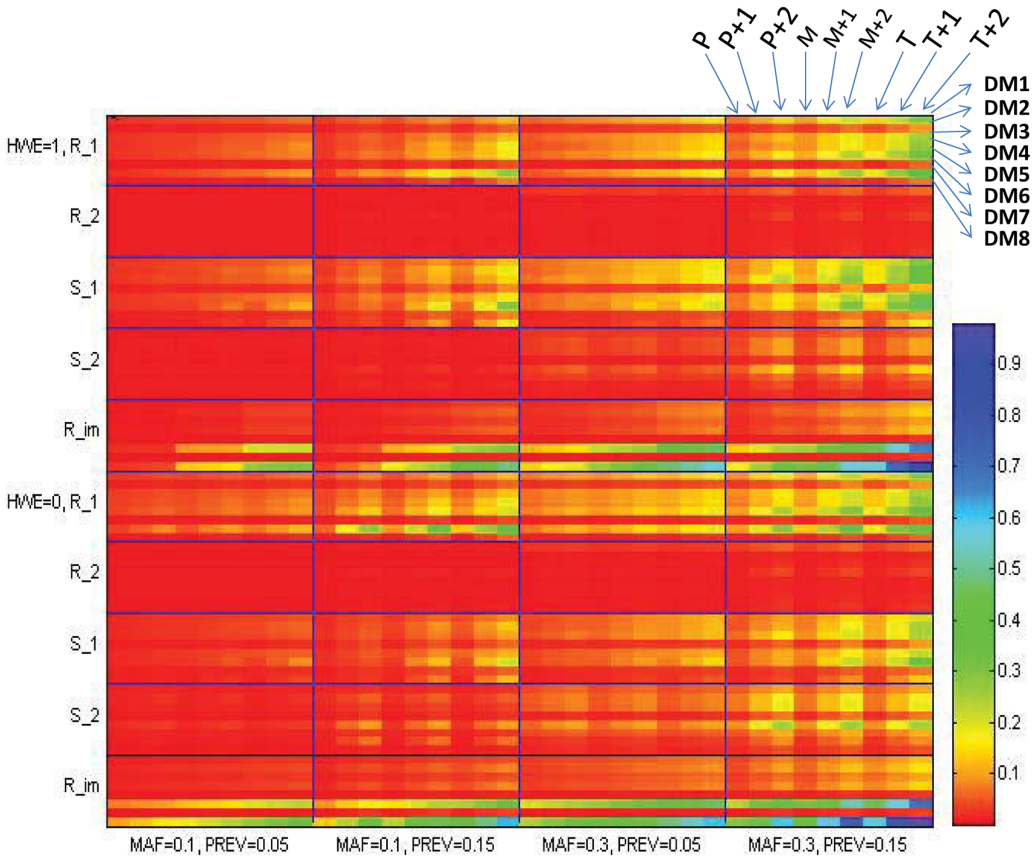


Figure 2. Information content per family for parameter estimation from the nine study designs (data types): $\{P, P+1, P+2, M, M+1, M+2, T, T+1, T+2\}$. Each of the four column blocks represent an MAF and PREV combination. Within each block, information from the nine data types are presented in the order 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_{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_{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-

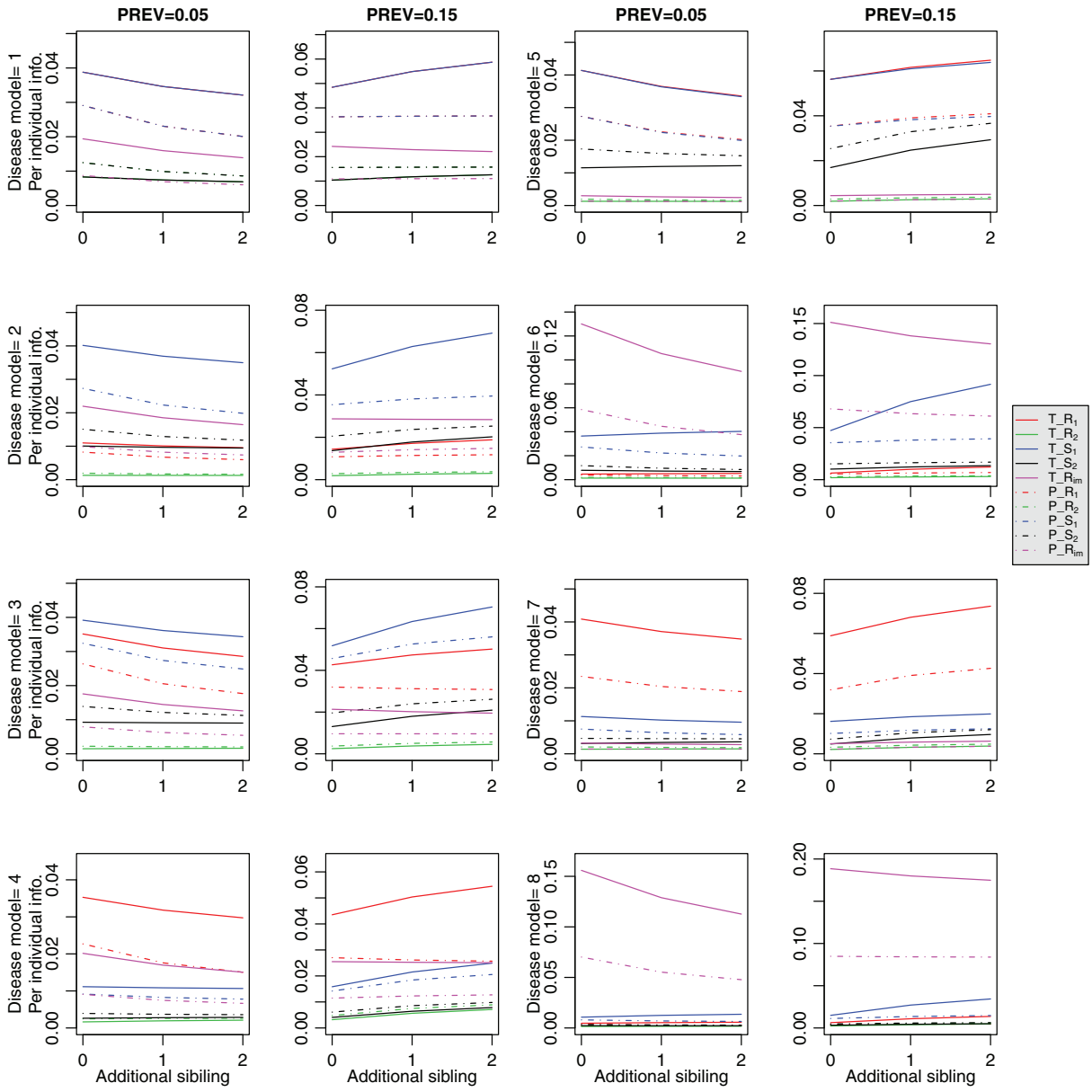


Figure 3. Information content per individual for eight disease models and two PREVs with MAF = 0.3 and HWE = 1. Each curve provides the information for estimating one of the five parameters, for a particular family type with 0, 1, or with 2 additional siblings. The study designs considered are the *T* 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_{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.

Supplementary Materials for "Optimum Study Design for Detecting Imprinting and Maternal Effects Based on Partial Likelihood"

Fangyuan Zhang, Abbas Khalili, Shili Lin

A.1 Technical Details Related to Theorem 1

Recall that LIME uses a multiplicative relative risk model for the disease prevalence:

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

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

The vector of parameters of interest is denoted by

$$\boldsymbol{\theta} = (\delta, R_1, R_2, R_{im}, S_1, S_2).$$

To make inference about $\boldsymbol{\theta}$, we use the partial log-likelihood

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

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

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

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 $eN^j_p = \sum_{(m,c)\neq(1,1)} n^j_{mc}$, and $esN^j_p = \sum_{(m,c)\neq(1,1)} sn^j_{mc}$, for $j = 0, 1$. Hence $(n_t + en_p)$ is the total number of independent 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.

The *maximum partial likelihood estimator* (MPLE) of $\boldsymbol{\theta}$ is denoted by

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

which is assumed to be obtained by solving the score-type equation

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

In the next section we study the theoretical properties of $\widehat{\boldsymbol{\theta}}_n$, as the effective sample size $n = n_t + en_p + sn_t + esn_p$ tends to infinity. We should note that here when $n \to \infty$, each of the sample sizes (n_t, en_p, sn_t, esn_p) also tends to infinity, at the same rate, such that

$$\frac{n_t}{n} \longrightarrow 1 \, , \; \frac{en_p}{n} \longrightarrow \; , \; \frac{sn_t}{n} \longrightarrow 1 \, , \; \frac{esn_p}{n} \longrightarrow 1.$$

Clearly, this is under the assumption that all the four sums $\sum$ are present in the partial log-likelihood $l_{par}(\boldsymbol{\theta})$ defined above. If any of the terms are not present (for example, in a trio setting where only $l_{t1}(\boldsymbol{\theta})$ exists), the theorem still holds and the proof is analogous.

A.1.1 Regularity Conditions

Let $\boldsymbol{\theta}_0$ be the true value of the parameter of interest. In what follows we denote

$$C_{r_n}(\boldsymbol{\theta}_0) = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^6 : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq r_n\}$$

as some neighborhood of $\boldsymbol{\theta}_0$, with radius r_n, where $r_n \to 0$, as n tends to infinity. Later on, we will see that this rate is $n^{-1/2}$. The regularity conditions are:

R1. The true value $\boldsymbol{\theta}_0$ of the parameter vector $\boldsymbol{\theta}$ is an interior point of the compact parameter space $\boldsymbol{\Theta}$.

R2. The cell probabilities $p_{mfc}(\boldsymbol{\theta})$, $p_{mc}(\boldsymbol{\theta})$, $q_{mfc}(\boldsymbol{\theta})$ and $q_{mc}(\boldsymbol{\theta})$ admit up to their third-order partial derivatives with respect to the elements of the parameter vector $\boldsymbol{\theta} = (\delta, R_1, R_2, R_{im}, S_1, S_2)$, for any $\boldsymbol{\theta} \in C_{r_n}(\boldsymbol{\theta}_0)$.

R3. The cell probabilities $p_{mfc}(\boldsymbol{\theta})$, $p_{mc}(\boldsymbol{\theta})$, $q_{mfc}(\boldsymbol{\theta})$ and $q_{mc}(\boldsymbol{\theta})$ are bounded away from the boundaries zero and one, and their partial derivatives $p'_{mfc}(\boldsymbol{\theta})$, $p'_{mc}(\boldsymbol{\theta})$, $q'_{mfc}(\boldsymbol{\theta})$ and $q'_{mc}(\boldsymbol{\theta})$ are bounded away from zero, for those $\boldsymbol{\theta} \in C_{r_n}(\boldsymbol{\theta}_0)$. Further, the partial derivatives of the cell probabilities, up to third order, are bounded by some constants, for any $\boldsymbol{\theta} \in C_{r_n}(\boldsymbol{\theta}_0)$.

R4. Identifiability: for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$, $p_{mfc}(\boldsymbol{\theta}_1) = p_{mfc}(\boldsymbol{\theta}_2)$, $q_{mfc}(\boldsymbol{\theta}_1) = q_{mfc}(\boldsymbol{\theta}_2)$, for all (m, f, c) combinations and all $p_{mc}(\boldsymbol{\theta}_1) = p_{mc}(\boldsymbol{\theta}_2)$, $q_{mc}(\boldsymbol{\theta}_1) = q_{mc}(\boldsymbol{\theta}_2)$, for all $(m, c) \neq (1, 1)$, imply that $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$.

R5. The information matrix

$$\boldsymbol{I}(\boldsymbol{\theta}) = -E\{l''_{\text{par}}(\boldsymbol{\theta})\} = -E\left\{ \frac{\partial^2 l_{\text{par}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \right\}$$

is positive definite for any $\boldsymbol{\theta} \in C_{r_n}(\boldsymbol{\theta}_0)$.

We adopt the line of proof provided in Chanda (1954) and Lindsay (1980) to our partial likelihood context.

A.1.2 Proof of Theorem 1

Proof of Part (i) of Theorem 1. For simplicity in notation, we denote the vector of parameters of interest as $\boldsymbol{\theta} = (\delta, R_1, R_2, R_{im}, S_1, S_2) = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6)$. By the regularity Condition R2, for the first part of the partial log-likelihood, $l_{t1}(\boldsymbol{\theta})$, representing proband triads, we have that

$$\frac{\partial l_{t1}(\boldsymbol{\theta})}{\partial \theta_j} = l'_{t1,j}(\boldsymbol{\theta}) = l'_{t1,j}(\boldsymbol{\theta}_0) + \sum_{k=1}^{6} l''_{t1,jk}(\boldsymbol{\theta}_0)(\theta_k - \theta_k^0) + \frac{1}{2} \sum_{l,k}^{6} l'''_{t1,jkl}(\widetilde{\boldsymbol{\theta}})(\theta_k - \theta_k^0)(\theta_l - \theta_l^0) \quad (1)$$

for $j = 1, 2, \ldots, 6$, where $\widetilde{\boldsymbol{\theta}}$ is between $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta} \in C_{r_n}(\boldsymbol{\theta}_0)$; $l''_{t1,jk}(\cdot)$ and $l'''_{t1,jkl}(\cdot)$ are the second and third-order partial derivatives of the function $l_{t1}(\cdot)$, respectively. For $j, k, l =$

$1, 2, 3, 4, 5, 6$, we have

$$
\begin{aligned}
l'_{t1,j}(\boldsymbol{\theta}) &= \sum_{m,f,c} \frac{\partial p_{mfc}(\boldsymbol{\theta})}{\partial \theta_j} \times \left\{ \frac{n^1_{mfc}}{p_{mfc}(\boldsymbol{\theta})} - \frac{n_{mfc} - n^1_{mfc}}{1 - p_{mfc}(\boldsymbol{\theta})} \right\} \\
l''_{t1,jk}(\boldsymbol{\theta}) &= \sum_{m,f,c} \frac{\partial^2 p_{mfc}(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k} \times \left\{ \frac{n^1_{mfc}}{p_{mfc}(\boldsymbol{\theta})} - \frac{n_{mfc} - n^1_{mfc}}{1 - p_{mfc}(\boldsymbol{\theta})} \right\} \\
&- \sum_{m,f,c} \frac{\partial p_{mfc}(\boldsymbol{\theta})}{\partial \theta_j} \times \frac{\partial p_{mfc}(\boldsymbol{\theta})}{\partial \theta_k} \times \left\{ \frac{n^1_{mfc}}{[p_{mfc}(\boldsymbol{\theta})]^2} + \frac{n_{mfc} - n^1_{mfc}}{[1 - p_{mfc}(\boldsymbol{\theta})]^2} \right\} \\
l'''_{t1,jkl}(\boldsymbol{\theta}) &= \sum_{m,f,c} \frac{\partial^3 p_{mfc}(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k \partial \theta_l} \times \left\{ \frac{n^1_{mfc}}{p_{mfc}(\boldsymbol{\theta})} - \frac{n_{mfc} - n^1_{mfc}}{1 - p_{mfc}(\boldsymbol{\theta})} \right\} \\
&- \sum_{m,f,c} \frac{\partial^2 p_{mfc}(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k} \times \frac{\partial p_{mfc}(\boldsymbol{\theta})}{\partial \theta_l} \times \left\{ \frac{n^1_{mfc}}{[p_{mfc}(\boldsymbol{\theta})]^2} + \frac{n_{mfc} - n^1_{mfc}}{[1 - p_{mfc}(\boldsymbol{\theta})]^2} \right\} \\
&- \sum_{m,f,c} \left[\frac{\partial^2 p_{mfc}(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_l} \times \frac{\partial p_{mfc}(\boldsymbol{\theta})}{\partial \theta_k} + \frac{\partial p_{mfc}(\boldsymbol{\theta})}{\partial \theta_j} \times \frac{\partial^2 p_{mfc}(\boldsymbol{\theta})}{\partial \theta_k \partial \theta_l} \right] \times \left\{ \frac{n^1_{mfc}}{[p_{mfc}(\boldsymbol{\theta})]^2} + \frac{n_{mfc} - n^1_{mfc}}{[1 - p_{mfc}(\boldsymbol{\theta})]^2} \right\} \\
&- \sum_{m,f,c} \frac{\partial p_{mfc}(\boldsymbol{\theta})}{\partial \theta_j} \times \frac{\partial p_{mfc}(\boldsymbol{\theta})}{\partial \theta_k} \times \frac{\partial p_{mfc}(\boldsymbol{\theta})}{\partial \theta_l} \left\{ \frac{-2n^1_{mfc}}{[p_{mfc}(\boldsymbol{\theta})]^3} + \frac{2(n_{mfc} - n^1_{mfc})}{[1 - p_{mfc}(\boldsymbol{\theta})]^3} \right\}
\end{aligned}
$$

for any $\boldsymbol{\theta} \in C_{r_n}(\boldsymbol{\theta}_0)$.

For every triad type (m, f, c), denote the ratio

$$r^1_{mfc} = \frac{n^1_{mfc}}{n_{mfc}}$$

where $n_{mfc} = n^0_{mfc} + n^1_{mfc}$. The form of the partial log-likelihood $l_{par}(\boldsymbol{\theta})$ suggests that, for each triad type (m, f, c) and conditional on n_{mfc}, we have $n^1_{mfc}|n_{mfc} \sim Binomial(n_{mfc}, p_{mfc}(\boldsymbol{\theta}))$. By using a double conditional expectation technique, we have that $E(r^1_{mfc}) = p_{mfc}(\boldsymbol{\theta})$, where $E(\cdot)$ is the expected value under the model with the parameter-vector value $\boldsymbol{\theta}$. Thus,

$$
\begin{aligned}
n^{-1} E\{l'_{t1,j}(\boldsymbol{\theta})\} &= 0 \\
-n^{-1} E\{l''_{t1,jk}(\boldsymbol{\theta})\} &= \sum_{m,f,c} \frac{\partial p_{mfc}(\boldsymbol{\theta})}{\partial \theta_j} \times \frac{\partial p_{mfc}(\boldsymbol{\theta})}{\partial \theta_k} \times \left\{ \frac{E(n_{mfc}/n)}{[p_{mfc}(\boldsymbol{\theta})][1 - p_{mfc}(\boldsymbol{\theta})]} \right\} = I_{t1,jk}(\boldsymbol{\theta})
\end{aligned}
$$

for any $\boldsymbol{\theta} \in C_{r_n}(\boldsymbol{\theta}_0)$.

Further, by the regularity condition R3, for any $\boldsymbol{\theta} \in C_{r_n}(\boldsymbol{\theta}_0)$,

$$
\begin{aligned}
n^{-1}|l'''_{t1,jkl}(\boldsymbol{\theta})| &\leq \sum_{m,f,c} 2\left|\frac{\partial^3 p_{mfc}(\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k\partial\theta_l}\right| + \sum_{m,f,c}\left|\frac{\partial^2 p_{mfc}(\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k}\times\frac{\partial p_{mfc}(\boldsymbol{\theta})}{\partial\theta_l}\right|\times\left\{\frac{(n_{mfc}/n)}{[p_{mfc}(\boldsymbol{\theta})][1-p_{mfc}(\boldsymbol{\theta})]}\right\} \\
&+ \sum_{m,f,c}\left|\frac{\partial^2 p_{mfc}(\boldsymbol{\theta})}{\partial\theta_j\partial\theta_l}\times\frac{\partial p_{mfc}(\boldsymbol{\theta})}{\partial\theta_k}+\frac{\partial p_{mfc}(\boldsymbol{\theta})}{\partial\theta_j}\times\frac{\partial^2 p_{mfc}(\boldsymbol{\theta})}{\partial\theta_k\partial\theta_l}\right|\times\left\{\frac{(n_{mfc}/n)}{[p_{mfc}(\boldsymbol{\theta})][1-p_{mfc}(\boldsymbol{\theta})]}\right\} \\
&+ 2\sum_{m,f,c}\left|\frac{\partial p_{mfc}(\boldsymbol{\theta})}{\partial\theta_j}\times\frac{\partial p_{mfc}(\boldsymbol{\theta})}{\partial\theta_k}\times\frac{\partial p_{mfc}(\boldsymbol{\theta})}{\partial\theta_l}\right|\left\{\frac{(n_{mfc}/n)}{[p_{mfc}(\boldsymbol{\theta})]^2}+\frac{(n_{mfc}/n)}{[1-p_{mfc}(\boldsymbol{\theta})]^2}\right\} \\
&= O_p(1),
\end{aligned}
$$

which implies that $l'''_{t1,jkl}(\boldsymbol{\theta}) = O_p(n)$, for any $\boldsymbol{\theta}\in C_{r_n}(\boldsymbol{\theta}_0)$.

On the other hand, by the law of large numbers, we have that

$$
r^1_{mfc} = \frac{n^1_{mfc}}{n_{mfc}} \xrightarrow{w.p.o} p_{mfc}(\boldsymbol{\theta}_0)\ ,\ \ \frac{n_{mfc}}{n}\xrightarrow{w.p.o} E\left(\frac{n_{mfc}}{n}\right) = B_{mfc} \tag{2}
$$

for some constant $0\leq B_{mfc}<1$, as $n\to\infty$, where w.p.o stands for with probability tending to one. Thus, using (2), as $n\to\infty$, we have

$$
l'_{t1,j}(\boldsymbol{\theta}_0)/n \xrightarrow{w.p.o} 0\ ,\ \ -l''_{t1,jk}(\boldsymbol{\theta}_0)/n \xrightarrow{w.p.o} I_{t1,jk}(\boldsymbol{\theta}_0)\ ,\ l'''_{t1,jkl}(\boldsymbol{\theta}_0)/n = O_p(1). \tag{3}
$$

for $j,k,l = 1,2,\ldots,6$.

By similar arguments and under the regularity conditions R1-R5, for the remaining three terms of the partial log-likelihood, we have that

$$
\begin{aligned}
n^{-1}E\{l'_{p1,j}(\boldsymbol{\theta})\} &= 0 \\
-n^{-1}E\{l''_{p1,jk}(\boldsymbol{\theta})\} &= \sum_{(m,c)\neq(1,1)}\frac{\partial p_{mc}(\boldsymbol{\theta})}{\partial\theta_j}\times\frac{\partial p_{mc}(\boldsymbol{\theta})}{\partial\theta_k}\times\left\{\frac{E(n_{mc}/n)}{[p_{mc}(\boldsymbol{\theta})][1-p_{mc}(\boldsymbol{\theta})]}\right\} = I_{p1,jk}(\boldsymbol{\theta}) \\
n^{-1}\{l'''_{p1,jkl}(\boldsymbol{\theta})\} &= O_p(1)\ \text{ as }\ n\to\infty,
\end{aligned}
$$

$$
\begin{aligned}
n^{-1}E\{l'_{t2,j}(\boldsymbol{\theta})\} &= 0 \\
-n^{-1}E\{l''_{t2,jk}(\boldsymbol{\theta})\} &= \sum_{(m,f,c)}\frac{\partial q_{mfc}(\boldsymbol{\theta})}{\partial\theta_j}\times\frac{\partial q_{mfc}(\boldsymbol{\theta})}{\partial\theta_k}\times\left\{\frac{E(sn_{mfc}/n)}{[q_{mfc}(\boldsymbol{\theta})][1-q_{mfc}(\boldsymbol{\theta})]}\right\} = I_{t2,jk}(\boldsymbol{\theta}) \\
n^{-1}\{l'''_{t2,jkl}(\boldsymbol{\theta})\} &= O_p(1)\ \text{ as }\ n\to\infty,
\end{aligned}
$$

$$
\begin{aligned}
n^{-1}E\{l'_{p2,j}(\boldsymbol{\theta})\} &= 0 \\
n^{-1}E\{l''_{p2,jk}(\boldsymbol{\theta})\} &= \sum_{(m,c)\neq(1,1)}\frac{\partial q_{mc}(\boldsymbol{\theta})}{\partial\theta_j}\times\frac{\partial q_{mc}(\boldsymbol{\theta})}{\partial\theta_k}\times\left\{\frac{E(sn_{mc}/n)}{[q_{mc}(\boldsymbol{\theta})][1-q_{mc}(\boldsymbol{\theta})]}\right\} = I_{p2,jk}(\boldsymbol{\theta}) \\
n^{-1}\{l'''_{p2,jkl}(\boldsymbol{\theta})\} &= O_p(1)\ \text{ as }\ n\to\infty.
\end{aligned}
$$

Thus, similar to (3), as $n \to \infty$, we have that

$$l'_{p1,j}(\boldsymbol{\theta}_0)/n \xrightarrow{w.p.o} 0\ ,\ -l''_{p1,jk}(\boldsymbol{\theta}_0)/n \xrightarrow{w.p.o} I_{p1,jk}(\boldsymbol{\theta}_0)\ ,\ l'''_{p1,jkl}(\boldsymbol{\theta}_0)/n = O_p(1),$$

$$l'_{t2,j}(\boldsymbol{\theta}_0)/n \xrightarrow{w.p.o} 0\ ,\ -l''_{t2,jk}(\boldsymbol{\theta}_0)/n \xrightarrow{w.p.o} I_{t2jk}(\boldsymbol{\theta}_0)\ ,\ l'''_{t2,jkl}(\boldsymbol{\theta}_0)/n = O_p(1),$$

$$l'_{p2,j}(\boldsymbol{\theta}_0)/n \xrightarrow{w.p.o} 0\ ,\ -l''_{p2,jk}(\boldsymbol{\theta}_0)/n \xrightarrow{w.p.o} I_{p2,jk}(\boldsymbol{\theta}_0)\ ,\ l'''_{p2,jkl}(\boldsymbol{\theta}_0)/n = O_p(1)$$

for $j, k, l = 1, 2, \ldots, 6$, where similar to (2),

$$E\left(\frac{n_{mc}}{n}\right) = B_{mc}\ ,\ E\left(\frac{sn_{mfc}}{n}\right) = C_{mfc}\ ,\ E\left(\frac{sn_{mc}}{n}\right) = C_{mc}. \tag{4}$$

for some constants $0 \leq B_{mc} < 1$, $0 \leq C_{mfc} < 1$ and $0 \leq C_{mc} < 1$.

Using the above results, we have that

$$l'_{par}(\boldsymbol{\theta}_0)/n \xrightarrow{w.p.o} 0\ ,\ -l''_{par}(\boldsymbol{\theta}_0)/n \xrightarrow{w.p.o} \boldsymbol{I}(\boldsymbol{\theta}_0)\ ,\ l'''_{par}(\boldsymbol{\theta}_0)/n = O_p(1) \tag{5}$$

as $n \to \infty$. Here $\boldsymbol{I}(\boldsymbol{\theta}_0)$ is a 6×6 information matrix constructed based on the $\{I_{t1,jk}(\boldsymbol{\theta})$, $I_{p1,jk}(\boldsymbol{\theta})$, $I_{t2,jk}(\boldsymbol{\theta})$, $I_{p2,jk}(\boldsymbol{\theta})\}$, for $j, k = 1, 2, \ldots, 6$.

Thus consider the score-type equation divided by the total sample size n, which leads to the equations

$$n^{-1}\sum_{k=1}^{6} l''_{par,jk}(\boldsymbol{\theta}_0)(\theta_k - \theta_k^0) = -n^{-1}l'_{par,j}(\boldsymbol{\theta}_0) - \frac{1}{2}n^{-1}\sum_{l,k=1}^{6} l'''_{par,jkl}(\widetilde{\boldsymbol{\theta}})(\theta_k - \theta_k^0)(\theta_l - \theta_l^0)$$

for $j = 1, \ldots, 6$. By expanding the summation on the left hand side and re-writing with respect to each $\theta_k - \theta_k^0$, we have that

$$\theta_k - \theta_k^0 = \sum_{j=1}^{6}[\frac{-1}{n}l'_{par,j}(\boldsymbol{\theta}_0)] \times l^*_{par,jk}(\boldsymbol{\theta}_0) - \frac{1}{2}\sum_{l,r=1}^{6}\left[(\theta_r - \theta_r^0)(\theta_l - \theta_l^0)\left(\sum_{j=1}^{6}[\frac{1}{n}l'''_{par,jrl}(\widetilde{\boldsymbol{\theta}})] \times l^*_{par,jk}(\boldsymbol{\theta}_0)\right)\right] \tag{6}$$

for $k = 1, \ldots, 6$, where $l^*_{par,jk}(\boldsymbol{\theta}_0)$ are the elements of the inverse matrix $\Big(l''_{par,jk}(\boldsymbol{\theta}_0)/n; j, k = 1, \ldots, 6\Big)^{-1}$. By (5), the first term on the right hand side of the above equations tends to zero, as $n \to \infty$. This implies that the equations in (6) have at least one solution, in terms of $\theta_k - \theta_k^0$, that satisfies

$$\hat{\theta}_k - \theta_k^0 \longrightarrow^p 0\ \ ;\ \ k = 1, \ldots, 6,$$

as $n \to \infty$. Thus, there exists a solution, say, $\widehat{\boldsymbol{\theta}}_n$ of the score-type equation $l'_{par}(\boldsymbol{\theta}) = \mathbf{0}$ such that $\widehat{\boldsymbol{\theta}}_n \longrightarrow^p \boldsymbol{\theta}_0$, as $n \to \infty$.

Next stage is to prove the uniqueness of such consistent estimator. Under the regularity conditions R1-R5, and consistency of $\widehat{\boldsymbol{\theta}}_n$, we have that

$$\frac{1}{n} l''_{\text{par}}(\widehat{\boldsymbol{\theta}}_n) + \boldsymbol{I}(\boldsymbol{\theta}_0) = o_p(1) \tag{7}$$

as n tends to ∞, where $\boldsymbol{I}(\boldsymbol{\theta}_0)$ is the positive definite information matrix. Let us assume that there exist two such consistent estimators, say, $\widehat{\boldsymbol{\theta}}_{1n}$ and $\widehat{\boldsymbol{\theta}}_{2n}$ of $\boldsymbol{\theta}_0$ that are the solutions of the score-type equation

$$l'_{\text{par}}(\boldsymbol{\theta}) = 0.$$

By the extension of Rolle's theorem to multivariate case, there exists a point $\widetilde{\boldsymbol{\theta}}_n$ laying inside a hyper-cell with the vector $\widehat{\boldsymbol{\theta}}_{1n} - \widehat{\boldsymbol{\theta}}_{2n}$ as its diagonal, such that

$$l''_{\text{par}}(\widetilde{\boldsymbol{\theta}}_n) = 0 \tag{8}$$

On the other hand, since $\widehat{\boldsymbol{\theta}}_{1n}$ and $\widehat{\boldsymbol{\theta}}_{2n}$ are consistent estimators, so is $\widetilde{\boldsymbol{\theta}}_n$ and it must satisfy (7). But clearly (7) and (8) contradict. This implies that the consistent estimator $\widehat{\boldsymbol{\theta}}_n$ is unique. This completes the proof of Part(i). ♠

The result of Lemma 1 below is used for proving Part (ii) of Theorem 1.

Lemma 1 *Under the regularity conditions R1-R5, we have that*

$$\frac{l'_{par}(\boldsymbol{\theta}_0)}{\sqrt{n}} \longrightarrow^d N(\mathbf{0}, \boldsymbol{I}(\boldsymbol{\theta}_0))$$

as $n \to \infty$.

Proof of Lemma 1. Consider the partial-score function

$$\begin{aligned}
\left.\frac{\partial l_{par}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = l'_{par}(\boldsymbol{\theta}_0) &= l'_{t1}(\boldsymbol{\theta}_0) + l'_{p1}(\boldsymbol{\theta}_0) + l'_{t2}(\boldsymbol{\theta}_0) + l'_{p2}(\boldsymbol{\theta}_0) \\
&= \sum_{m,f,c} \frac{n_{mfc} \times p'_{mfc}(\boldsymbol{\theta}_0)}{p_{mfc}(\boldsymbol{\theta}_0)[1 - p_{mfc}(\boldsymbol{\theta}_0)]} \times [r^1_{mfc} - p_{mfc}(\boldsymbol{\theta}_0)] \\
&+ \sum_{(m,c)\neq(1,1)} \frac{n_{mc} \times p'_{mc}(\boldsymbol{\theta}_0)}{p_{mc}(\boldsymbol{\theta}_0)[1 - p_{mc}(\boldsymbol{\theta}_0)]} \times [r^1_{mc} - p_{mc}(\boldsymbol{\theta}_0)] \\
&+ \sum_{m,f,c} \frac{sn_{mfc} \times q'_{mfc}(\boldsymbol{\theta}_0)}{q_{mfc}(\boldsymbol{\theta}_0)[1 - q_{mfc}(\boldsymbol{\theta}_0)]} \times [s^1_{mfc} - q_{mfc}(\boldsymbol{\theta}_0)] \\
&+ \sum_{(m,c)\neq(1,1)} \frac{sn_{mc} \times q'_{mc}(\boldsymbol{\theta}_0)}{q_{mc}(\boldsymbol{\theta}_0)[1 - q_{mc}(\boldsymbol{\theta}_0)]} \times [s^1_{mc} - q_{mc}(\boldsymbol{\theta}_0)],
\end{aligned}$$

where $p'_{mfc}(\boldsymbol{\theta}_0), p'_{mc}(\boldsymbol{\theta}_0), q'_{mfc}(\boldsymbol{\theta}_0)$ and $q'_{mc}(\boldsymbol{\theta}_0)$ are the 6-dimensional vectors of the partial derivatives of the cell probabilities $p_{mfc}(\boldsymbol{\theta}), p_{mc}(\boldsymbol{\theta}), q_{mfc}(\boldsymbol{\theta})$ and $q_{mc}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, which are evaluated at the true $\boldsymbol{\theta}_0$. Also,

$$r^1_{mfc} = \frac{n^1_{mfc}}{n_{mfc}} \ , \ r^1_{mc} = \frac{n^1_{mc}}{n_{mc}} \ , \ s^1_{mfc} = \frac{sn^1_{mfc}}{sn_{mfc}} \ , \ s^1_{mc} = \frac{sn^1_{mc}}{sn_{mc}},$$

are the ratios of the number of cases among: proband (m, f, c) triads, proband (m, c) pairs, additional (m, f, c) sibling triads, and additional (m, c) sibling pairs, respectively.

We first try to find the limiting distribution of $l'_{t1}(\boldsymbol{\theta}_0)/\sqrt{n}$, as $n \to \infty$. We have that

$$\frac{l'_{t1}(\boldsymbol{\theta}_0)}{\sqrt{n}} = \sum_{m,f,c} \frac{p'_{mfc}(\boldsymbol{\theta}_0)}{p_{mfc}(\boldsymbol{\theta}_0)[1 - p_{mfc}(\boldsymbol{\theta}_0)]} \times \sqrt{\frac{n_{mfc}}{n}} \times \sqrt{n_{mfc}} \, [r^1_{mfc} - p_{mfc}(\boldsymbol{\theta}_0)]$$

In what follows we use the Wald device. For any non-zero vector $\boldsymbol{v} \in \mathbb{R}^6$,

$$w_n(\boldsymbol{\theta}_0) = \frac{\boldsymbol{v}^\top l'_{t1}(\boldsymbol{\theta}_0)}{\sqrt{n}} = \sum_{m,f,c} \frac{u_{mfc}(\boldsymbol{\theta}_0)}{p_{mfc}(\boldsymbol{\theta}_0)[1 - p_{mfc}(\boldsymbol{\theta}_0)]} \times \sqrt{\frac{n_{mfc}}{n}} \times \sqrt{n_{mfc}} \, [r^1_{mfc} - p_{mfc}(\boldsymbol{\theta}_0)]$$

where $u_{mfc}(\boldsymbol{\theta}_0) = \boldsymbol{v}^\top p'_{mfc}(\boldsymbol{\theta}_0)$ is a scalar. Note that conditional on the n_{mfc}'s, the ratios r^1_{mfc}'s are independent, each having the conditional asymptotic distribution

$$\sqrt{n_{mfc}} \, [r^1_{mfc} - p_{mfc}(\boldsymbol{\theta}_0)] \longrightarrow^d N(0, p_{mfc}(\boldsymbol{\theta}_0)(1 - p_{mfc}(\boldsymbol{\theta}_0))$$

as $n \to \infty$. Since n_{mfc}'s are following a multinomial distribution, say, with the joint probability mass function $g(n_{mfc}; m, f, c)$, then

$$F_n(w) = P(w_n(\boldsymbol{\theta}_0) \le w) = \sum_{\{m,f,c:n_{mfc}=0\}}^{n_t} P(w_n(\boldsymbol{\theta}_0) \le w | n_{mfc}, m, f, c) \; g(n_{mfc}; m, f, c).$$

On the other hand, as $n \to \infty$, since $n_{mfc}/n \xrightarrow{w.p.o} E(n_{mfc}/n) = B_{mfc}$, for some constant $0 \le B_{mfc} < 1$, then

$$(w_n(\boldsymbol{\theta}_0) | n_{mfc}, m, f, c) \longrightarrow^d N(0, \sigma^2(\boldsymbol{\theta}_0))$$

where

$$\sigma^2(\boldsymbol{\theta}_0) = \sum_{m,f,c} \frac{u^2_{mfc}(\boldsymbol{\theta}_0) \times B_{mfc}}{p_{mfc}(\boldsymbol{\theta}_0)(1 - p_{mfc}(\boldsymbol{\theta}_0))}.$$

Therefore, for $w \in \mathbb{R}$, as $n \to \infty$,

$$F_n(w) \longrightarrow \frac{1}{\sigma(\boldsymbol{\theta}_0)} \Phi\left(\frac{w}{\sigma(\boldsymbol{\theta}_0)}\right)$$

where $\Phi(\cdot)$ is the distribution function of the standard normal. This implies that

$$w_n(\boldsymbol{\theta}_0) \longrightarrow^d N(0, \sigma^2(\boldsymbol{\theta}_0))$$

as $n \to \infty$. Hence,

$$\frac{l'_{t1}(\boldsymbol{\theta}_0)}{\sqrt{n}} \longrightarrow^d N\left(\mathbf{0}, \sum_{m,f,c} \frac{[p'_{mfc}(\boldsymbol{\theta}_0)][p'_{mfc}(\boldsymbol{\theta}_0)]^\top \times B_{mfc}}{p_{mfc}(\boldsymbol{\theta}_0)(1 - p_{mfc}(\boldsymbol{\theta}_0))}\right) \ , \ n \to \infty.$$

Similarly, we have

$$\begin{aligned}
\frac{l'_{p1}(\boldsymbol{\theta}_0)}{\sqrt{n}} &\longrightarrow^d N\left(\mathbf{0}, \sum_{(m,c)\neq(1,1)} \frac{[p'_{mc}(\boldsymbol{\theta}_0)][p'_{mc}(\boldsymbol{\theta}_0)]^\top \times B_{mc}}{p_{mc}(\boldsymbol{\theta}_0)(1 - p_{mc}(\boldsymbol{\theta}_0))}\right), \\
\frac{l'_{t2}(\boldsymbol{\theta}_0)}{\sqrt{n}} &\longrightarrow^d N\left(\mathbf{0}, \sum_{m,f,c} \frac{[q'_{mfc}(\boldsymbol{\theta}_0)][q'_{mfc}(\boldsymbol{\theta}_0)]^\top \times C_{mfc}}{q_{mfc}(\boldsymbol{\theta}_0)(1 - q_{mfc}(\boldsymbol{\theta}_0))}\right), \\
\frac{l'_{p2}(\boldsymbol{\theta}_0)}{\sqrt{n}} &\longrightarrow^d N\left(\mathbf{0}, \sum_{m,f,c} \frac{[q'_{mc}(\boldsymbol{\theta}_0)][q'_{mc}(\boldsymbol{\theta}_0)]^\top \times C_{mc}}{q_{mc}(\boldsymbol{\theta}_0)(1 - q_{mc}(\boldsymbol{\theta}_0))}\right),
\end{aligned}$$

for some constants $0 \leq B_{mc} < 1, 0 \leq C_{mfc} < 1$ and $0 \leq C_{mc} < 1$, introduced previously, such that, as $n \to \infty$,

$$\frac{n_{mc}}{n} \xrightarrow{w.p.o} E\left(\frac{n_{mc}}{n}\right) = B_{mc} \ , \ \frac{sn_{mfc}}{n} \xrightarrow{w.p.o} E\left(\frac{sn_{mfc}}{n}\right) = C_{mfc} \ , \ \frac{sn_{mc}}{n} \xrightarrow{w.p.o} E\left(\frac{sn_{mc}}{n}\right) = C_{mc}.$$

Note that the ratios r^1_{mfc} , r^1_{mc} , s^1_{mfc} and s^1_{mc} are independent under the partial likelihood formulation. Thus, as the effective sample size $n = n_t + en_p + sn_t + esn_p$ tends to infinity, we have

$$\frac{l'_{par}(\boldsymbol{\theta}_0)}{\sqrt{n}} = \frac{l'_{t1}(\boldsymbol{\theta}_0)}{\sqrt{n}} + \frac{l'_{p1}(\boldsymbol{\theta}_0)}{\sqrt{n}} + \frac{l'_{t2}(\boldsymbol{\theta}_0)}{\sqrt{n}} + \frac{l'_{p2}(\boldsymbol{\theta}_0)}{\sqrt{n}} \longrightarrow^d N\left(\mathbf{0}, \boldsymbol{I}(\boldsymbol{\theta}_0)\right)$$

where $\boldsymbol{I}(\boldsymbol{\theta}_0) = \boldsymbol{I}_{t1}(\boldsymbol{\theta}_0) + \boldsymbol{I}_{p1}(\boldsymbol{\theta}_0) + \boldsymbol{I}_{t2}(\boldsymbol{\theta}_0) + I_{p2}(\boldsymbol{\theta}_0)$, and

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

are 6×6-dimensional positive definite information matrices.

Hence, as $n \to \infty$, we have that

$$\frac{l'_{par}(\boldsymbol{\theta}_0)}{\sqrt{n}} \longrightarrow^d N(\mathbf{0}, \boldsymbol{I}(\boldsymbol{\theta}_0)). \tag{9}$$

This completes the proof of Lemma 1. ♠

Proof of Part (ii) of Theorem 1. Let $\widehat{\boldsymbol{\theta}}_n$ be the MPLE, which satisfies the score-type equation

$$l'_{par}(\widehat{\boldsymbol{\theta}}_n) = 0.$$

By the regularity conditions R1-R5, we have that

$$\begin{aligned} \mathbf{0} &= \frac{1}{n} l'_{par}(\boldsymbol{\theta}_0) + \frac{1}{n} l''_{par}(\boldsymbol{\theta}_0)(1 + o_p(1)) \times (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \\ &= \frac{1}{n} l'_{par}(\boldsymbol{\theta}_0) + \left[\frac{1}{n} l''_{par}(\boldsymbol{\theta}_0) + \boldsymbol{I}(\boldsymbol{\theta}_0) - \boldsymbol{I}(\boldsymbol{\theta}_0)\right](1 + o_p(1)) \times (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \end{aligned}$$

where by (5) $l''_{par}(\boldsymbol{\theta}_0)/n + \boldsymbol{I}(\boldsymbol{\theta}_0) = o_p(1)$. Therefore, by the result of Lemma 1,

$$\sqrt{n}\,(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \boldsymbol{I}^{-1}(\boldsymbol{\theta}_0) \times \frac{l'_{par}(\boldsymbol{\theta}_0)}{\sqrt{n}} \longrightarrow^d N(0, \boldsymbol{I}^{-1}(\boldsymbol{\theta}_0)),$$

as $n \to \infty$. This completes the proof of Part(ii) of Theorem 1. ♠

A.1.3 Calculation of the Constants in the Information Matrix $\boldsymbol{I}(\boldsymbol{\theta}_0)$

In this section we provide an example where we calculate the four constants B_{mfc}, B_{mc}, C_{mfc} and C_{mc} in the information matrix $\boldsymbol{I}(\boldsymbol{\theta}_0)$. First note that,

$$\sum_{m,f,c} B_{mfc} + \sum_{(m,c)\neq(1,1)} B_{mc} + \sum_{m,f,c} C_{mfc} + \sum_{(m,c)\neq(1,1)} C_{mc} = 1.$$

In our example we consider mix families each with k additional siblings. Denote υ as the rate of missing father's genotype. Further, denote r_t^1 and r_t^0 as the proportion of affected and unaffected probands in complete families, respectively, r_p^1 and r_p^0 as the proportion of affected and unaffected probands in incomplete families, respectively. These are population level proportions. The number of independent nuclear families, i.e. the sample size, is N.

In the partial log-likelihood $l_{par}(\boldsymbol{\theta})$, the effective sample size is n, as also demonstrated in the main paper. Based on the above information, the expected effective sample size, which

we also, by the abuse of notation, represent it by n, is given by

$$
\begin{aligned}
& n = (k+1)N(1-\upsilon) + N\upsilon r_p^1(1 - P(M=1, C=1|D=1)) + N\upsilon r_p^0(1 - P(M=1, C=1|D=0)) \\
+ \quad & kN\upsilon r_p^1(1 - P(M=1, C_2=1|D_1=1)) + kN\upsilon r_p^0(1 - P(M=1, C_2=1|D_1=0)) \\
= \quad & (k+1)N(1-\upsilon) + N\upsilon r_p^1(1 - \frac{q_{11}P(M=1,C=1)}{P(D=1)}) + N\upsilon r_p^0(1 - \frac{(1-q_{11})P(M=1,C=1)}{P(D=0)}) \\
+ \quad & kN\upsilon r_p^1(1 - \frac{\sum_f \sum_{c*} P(M=1,F=f,C=c*)P(D=1|M=1,F=f,C=c*)P(C=1|M=1,F=f)}{P(D=1)}) \\
+ \quad & kN\upsilon r_p^0(1 - \frac{\sum_f \sum_{c*} P(M=1,F=f,C=c*)P(D=0|M=1,F=f,C=c*)P(C=1|M=1,F=f)}{P(D=0)}) \\
= \quad & N \times \Bigg\{ (k+1)(1-\upsilon) + \upsilon r_p^1 \Bigg(1 - \frac{q_{11}P(M=1,C=1)}{P(D=1)} \Bigg) + \upsilon r_p^0 \Bigg(1 - \frac{(1-q_{11})P(M=1,C=1)}{P(D=0)} \Bigg) \\
+ \quad & k\upsilon r_p^1 \Bigg(1 - \frac{\sum_f \sum_{c*} P(M=1,F=f,C=c*)P(D=1|M=1,F=f,C=c*)P(C=1|M=1,F=f)}{P(D=1)} \Bigg) \\
+ \quad & k\upsilon r_p^0 \Bigg(1 - \frac{\sum_f \sum_{c*} P(M=1,F=f,C=c*)P(D=0|M=1,F=f,C=c*)P(C=1|M=1,F=f)}{P(D=0)} \Bigg) \Bigg\} \\
= \quad & Nh
\end{aligned}
$$

where h is the expression within the curly brackets $\{\cdot\}$, and it is the ratio of the expected effective sample size n and total family count N, and is also independent of the total family count N.

In what follows, we use the relation $n = Nh$ to calculate the four constants B_{mfc}, B_{mc}, C_{mfc} and C_{mc}. Note that, from the proof of Theorem 1, we have that

$$
B_{mfc} = E\left(\frac{n_{mfc}}{n}\right) \ , \ B_{mc} = E\left(\frac{n_{mc}}{n}\right) \ , \ C_{mfc} = E\left(\frac{sn_{mfc}}{n}\right) \ , \ C_{mc} = E\left(\frac{sn_{mc}}{n}\right).
$$

Thus,

$$
\begin{aligned}
B_{mfc} &= \frac{N(1-\upsilon)r_t^1 P(M=m,F=f,C_1=c|D_1=1) + N(1-\upsilon)r_t^0 P(M=m,F=f,C_1=c|D_1=0)}{Nh} \\
&= \frac{(1-\upsilon)r_t^1 q_{mfc} P(M=m,F=f,C_1=c)}{hP(D_1=1)} + \frac{(1-\upsilon)r_t^0(1-q_{mfc})P(M=m,F=f,C_1=c)}{hP(D_1=0)},
\end{aligned}
$$

$$
\begin{aligned}
B_{mc} &= \frac{N\upsilon r_p^1 P(M=m,C_1=c|D_1=1) + N\upsilon r_p^0 P(M=m,C_1=c|D_1=0)}{Nh} \\
&= \frac{\upsilon r_p^1 q_{mc} P(M=m,C_1=c)}{hP(D_1=1)} + \frac{\upsilon r_p^0(1-q_{mc})P(M=m,C_1=c)}{hP(D_1=0)},
\end{aligned}
$$

$$
\begin{aligned}
C_{mfc} &= \frac{kN(1-\upsilon)r_t^1 P(M=m,F=f,C_2=c|D_1=1)+kN(1-\upsilon)r_t^0 P(M=m,F=f,C_2=c|D_1=0)}{Nh} \\
&= \frac{k(1-\upsilon)r_t^1 P(M=m,F=f,C_2=c)\sum_{c^*}P(C_1=c^*|M=m,F=f)q_{mfc^*}}{hP(D_1=1)} \\
&+ \frac{k(1-\upsilon)r_t^0 P(M=m,F=f,C_2=c)\sum_{c^*}P(C_1=c^*|M=m,F=f)(1-q_{mfc^*})}{hP(D_1=0)},
\end{aligned}
$$

and

$$
\begin{aligned}
C_{mc} &= \frac{kN\upsilon r_p^1 P(M=m,C_2=c|D_1=1)+kN\upsilon r_p^0 P(M=m,C_2=c|D_1=0)}{Nh} \\
&= \frac{k\upsilon r_p^1 \sum_f P(M=m,F=f,C_2=c)\sum_{c^*}P(C_1=c^*|M=m,F=f)q_{mfc^*}}{hP(D_1=1)} \\
&+ \frac{k\upsilon r_p^0 \sum_f P(M=m,F=f,C_2=c)\sum_{c^*}P(C_1=c^*|M=m,F=f)(1-q_{mfc^*})}{hP(D_1=0)}.
\end{aligned}
$$

Note that these constants are independent of the total family count (sample size) N.

A.2: Calculation of per family and per individual information content

A.2.1 Per family information content

The Fisher information matrix $\boldsymbol{I}(\boldsymbol{\theta}_0)$ given in Theorem 1 is the expected information *per effective family*. Denote $\boldsymbol{I}_{\text{fam}}(\boldsymbol{\theta}_0)$ as the expected *information per family*. Suppose there are k additional siblings in each family. As h is the ratio of effective sample size n and total count of families N, given in Subsection A.1.3, then

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

Thus, we have

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

In what follows we give three examples of data types, and the calculations of h and the four constants B_{mfc}, B_{mc}, C_{mfc} and C_{mc}.

Example 1. The data type is $P + k$, that is, all the independent nuclear families in the sample are of "pair type", each with k additional siblings. Then $B_{mfc} = C_{mfc} = 0$, and the proportion of families with father's genotype missing is $\upsilon = 1$. Thus

$$\begin{aligned} h &= r_p^1\left(1 - \frac{q_{11}P(M=1,C=1)}{P(D=1)}\right) + r_p^0\left(1 - \frac{(1-q_{11})P(M=1,C=1)}{P(D=0)}\right) \\ &+ kr_p^1\left(1 - \frac{\sum_f \sum_{c*} P(M=1,F=f,C=c*)P(D=1|M=1,F=f,C=c*)P(C=1|M=1,F=f)}{P(D=1)}\right) \\ &+ kr_p^0\left(1 - \frac{\sum_f \sum_{c*} P(M=1,F=f,C=c*)P(D=0|M=1,F=f,C=c*)P(C=1|M=1,F=f)}{P(D=0)}\right). \end{aligned}$$

Further, we have

$$\begin{aligned} B_{mc} &= \frac{N\upsilon r_p^1 P(M=m,C_1=c|D_1=1) + N\upsilon r_p^0 P(M=m,C_1=c|D_1=0)}{Nh} \\ &= \frac{r_p^1 q_{mc} P(M=m,C_1=c)}{hP(D_1=1)} + \frac{r_p^0(1-q_{mc})P(M=m,C_1=c)}{hP(D_1=0)}, \end{aligned}$$

$$\begin{aligned} C_{mc} &= \frac{kN\upsilon r_p^1 P(M=m,C_2=c|D_1=1) + kN\upsilon r_p^0 P(M=m,C_2=c|D_1=0)}{Nh} \\ &= \frac{kr_p^1 \sum_f [P(M=m,F=f,C_2=c) \times \sum_{c^*} P(C_1=c^*|M=m,F=f) q_{mfc^*}]}{hP(D_1=1)} \\ &+ \frac{kr_p^0 \sum_f [P(M=m,F=f,C_2=c) \times \sum_{c^*} P(C_1=c^*|M=m,F=f)(1-q_{mfc^*})]}{hP(D_1=0)}. \end{aligned}$$

Example 2. Consider the data type $M + k$, that is, a mixture of "triad" and "pair" type independent nuclear families, each with k additional siblings. We have that,

$$\begin{aligned} h &= (k+1)(1-\upsilon) + \upsilon r_p^1(1 - P(M=1,C=1|D=1)) + \upsilon r_p^0(1 - P(M=1,C=1|D=0)) \\ &+ \upsilon r_p^1(1 - P(M=1,C_2=1|D_1=1)) + \upsilon r_p^0(1 - P(M=1,C_2=1|D_1=0)) \\ &= (k+1)(1-\upsilon) + \upsilon r_p^1\left(1 - \frac{q_{11}P(M=1,C=1)}{P(D=1)}\right) + \upsilon r_p^0\left(1 - \frac{(1-q_{11})P(M=1,C=1)}{P(D=0)}\right) \\ &+ \upsilon r_p^1\left(1 - \frac{\sum_f \sum_{c*} P(M=1,F=f,C=c*)P(D=1|M=1,F=f,C=c*)P(C=1|M=1,F=f)}{P(D=1)}\right) \\ &+ \upsilon r_p^0\left(1 - \frac{\sum_f \sum_{c*} P(M=1,F=f,C=c*)P(D=0|M=1,F=f,C=c*)P(C=1|M=1,F=f)}{P(D=0)}\right). \end{aligned}$$

Further, B_{mfc}, B_{mc}, C_{mfc}, and C_{mc} are the same as the corresponding formula given on pages 9 and 10.

Example 3. Here we consider the data type T, which implies that

$$C_{mfc} = B_{mc} = C_{mc} = 0 \ , \ \upsilon = 0.$$

In this case, there are equal numbers of total family count and effective sample size, i.e. $h = 1$. Thus,

$$\begin{aligned} B_{mfc} &= \frac{N(1-\upsilon) r_t^1 P(M=m, F=f, C_1=c|D_1=1) + N(1-\upsilon) r_t^0 P(M=m, F=f, C_1=c|D_1=0)}{Nh} \\ &= \frac{r_t^1 q_{mfc} P(M=m, F=f, C_1=c)}{P(D_1=1)} + \frac{r_t^0 (1-q_{mfc}) P(M=m, F=f, C_1=c)}{P(D_1=0)}. \end{aligned}$$

A.2.2 Per individual information content

Denote the information matrix per individual as $\boldsymbol{I}_{\text{ind}}(\boldsymbol{\theta}_0)$. Let n^* be the total number of individuals involved (including all the fathers, mothers and offsprings), and denote g as the ratio of the total individuals involved n^* to the total family counts N, i.e. $g = n^*/N$. For example, if the data type is $P + 2$, then $g = 4$. If the data type is $M + 1$, then $g = 4(1-\upsilon) + 3\upsilon = 4 - \upsilon$. If the data type is T, we have $g = 3$. Then

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

Suppose there are k additional siblings per independent nuclear family, then

$$n^* = (k+3)N(1-\upsilon) + (k+2)N\upsilon = Ng.$$

Thus, the information matrix per individual is given by

$$\begin{aligned} I_{\text{ind}}(\boldsymbol{\theta}_0) = & \sum_{m,f,c} \frac{[p'_{mfc}(\boldsymbol{\theta}_0)][p'_{mfc}(\boldsymbol{\theta}_0)]^\top \times B_{mfc} \times \frac{h}{g}}{p_{mfc}(\boldsymbol{\theta}_0)(1-p_{mfc}(\boldsymbol{\theta}_0))} + \sum_{(m,c)\neq(1,1)} \frac{[p'_{mc}(\boldsymbol{\theta}_0)][p'_{mc}(\boldsymbol{\theta}_0)]^\top \times B_{mc} \times \frac{h}{g}}{p_{mc}(\boldsymbol{\theta}_0)(1-p_{mc}(\boldsymbol{\theta}_0)} \\ + & \sum_{m,f,c} \frac{[q'_{mfc}(\boldsymbol{\theta}_0)][q'_{mfc}(\boldsymbol{\theta}_0)]^\top \times C_{mfc} \times \frac{h}{g}}{q_{mfc}(\boldsymbol{\theta}_0)(1-q_{mfc}(\boldsymbol{\theta}_0))} + \sum_{(m,c)\neq(1,1)} \frac{[q'_{mc}(\boldsymbol{\theta}_0)][q'_{mc}(\boldsymbol{\theta}_0)]^\top \times C_{mc} \times \frac{h}{g}}{q_{mc}(\boldsymbol{\theta}_0)(1-q_{mc}(\boldsymbol{\theta}_0))}. \end{aligned}$$

We now look at two data types, and calculate h, g and the information matrix:

For data type $T + k$, we have:

$$g_{T+k} = k + 3 \ , \ h_{T+k} = k + 1.$$

For data type $P+k$, we have $g_{P+k} = k+2$, and

$$\begin{aligned} h_{P+k} &= r_p^1\Big(1 - \frac{q_{11}P(M=1,C=1)}{P(D=1)}) + r_p^0(1 - \frac{(1-q_{11})P(M=1,C=1)}{P(D=0)}\Big) \\ &+ kr_p^1\Big(1 - \frac{\sum_f \sum_{c*} P(M=1,F=f,C=c*)P(D=1|M=1,F=f,C=c*)P(C=1|M=1,F=f)}{P(D=1)} \\ &+ kr_p^0\Big(1 - \frac{\sum_f \sum_{c*} P(M=1,F=f,C=c*)P(D=0|M=1,F=f,C=c*)P(C=1|M=1,F=f)}{P(D=0)} \end{aligned}$$

Alternatively, we may re-write the information matrix per individual as,

$$\begin{aligned} \boldsymbol{I}_{\text{ind}}(\boldsymbol{\theta}_0) &= \Big(\sum_{m,f,c} \frac{[p'_{mfc}(\boldsymbol{\theta}_0)][p'_{mfc}(\boldsymbol{\theta}_0)]^\top \times B^t_{mfc}}{p_{mfc}(\boldsymbol{\theta}_0)(1-p_{mfc}(\boldsymbol{\theta}_0))} + \sum_{m,f,c} \frac{[q'_{mfc}(\boldsymbol{\theta}_0)][q'_{mfc}(\boldsymbol{\theta}_0)]^\top \times C^t_{mfc}}{q_{mfc}(\boldsymbol{\theta}_0)(1-q_{mfc}(\boldsymbol{\theta}_0))}\Big)\Big[\frac{h_{T+k}}{g_{T+k}}\, w_t\Big] \\ &+ \Big(\sum_{(m,c)\neq(1,1)} \frac{[p'_{mc}(\boldsymbol{\theta}_0)][p'_{mc}(\boldsymbol{\theta}_0)]^\top \times B^p_{mc}}{p_{mc}(\boldsymbol{\theta}_0)(1-p_{mc}(\boldsymbol{\theta}_0)} + \sum_{(m,c)\neq(1,1)} \frac{[q'_{mc}(\boldsymbol{\theta}_0)][q'_{mc}(\boldsymbol{\theta}_0)]^\top \times C^p_{mc}}{q_{mc}(\boldsymbol{\theta}_0)(1-q_{mc}(\boldsymbol{\theta}_0))}\Big)\Big[\frac{h_{P+k}}{g_{P+k}}\, w_p\Big], \end{aligned}$$

where $B^t_{mfc}, C^t_{mfc}, B^p_{mc}$, and C^p_{mc} are the limits, in probability, of $n_{mfc}/(n_t+sn_t)$, $sn_{mfc}/(n_t+sn_t)$, $n_{mc}/(en_p+esn_p)$, and $sn_{mc}/(en_p+esn_p)$, respectively. Further, w_t is the limit, in probability, of the proportion of the individuals in complete families, and w_p is the limit, in probability, of the proportion of the individuals in incomplete families, as the sample size increases. More specifically,

$$\begin{aligned} \frac{(k+3)N_t}{n^*} &\longrightarrow^p w_t = \frac{(k+3)(1-\upsilon)}{g}, \\ \frac{(k+2)N_p}{n^*} &\longrightarrow^p w_p = \frac{(k+2)\upsilon}{g}, \end{aligned}$$

as the sample size increases, where $w_t + w_p = 1$. Thus information per individual can be understood as either the information on average among all individuals or the weighted average of information per individual in complete families, and information per individual in incomplete families.

A.3: Verification of Asymptotic Properties

We use extensive simulations under a variety of disease models, scenarios and small to large sample sizes to verify the asymptotic properties of the maximum partial likelihood estimator $\widehat{\boldsymbol{\theta}}_n$ presented in Theorem 1 in the main paper. Specifically, we consider eight disease models as given in Table 1A in the main text. Under each model, we investigate eight combinations of scenarios concerning three factors (MAF, PREV, HWE; Table 1B in the main text). The

Supplementary Figure S1 is the QQ-norm plots for the relative difference between a parameter and its MPLE for data simulated under model 8 and scenario 5. The results are based on 500 replications with the number of nuclear families being $N = 200, 1000, 2000$ or 10000. Each family consists of two parents and three children, i.e., the proband (either diseased or normal) through which the family is ascertained, and two siblings. Each plot compares the quantiles of the relative difference with the quantiles of a normal distribution, for each of the model parameters. We can see that, as the sample size N increases, the relative differences get closer and closer to zero, implying possible consistency of the estimators. The figure also provides a clear visualization that the distribution of the relative difference approaches a normal distribution, as the QQ-norm plot follows the pattern of a straight line as N increases. The top segment of Table S1 provides the corresponding Kolmogorov-Smirnov (KS) distance and p-value by comparing the empirical distribution of the parameter estimators with normal distribution. These results indeed show that the empirical distribution is not distinguishable from a normal distribution based on a formal statistical test, corroborating the visual observations from Figure S1. KS distances and p-values as well as QQ-norm plots for the other disease models and scenarios (Supplementary Table S1 and Figures S2-S8 provide some of the results) lead to the same conclusion.

A.4: Sample Size Calculation

Based on the asymptotic variance given in Theorem 1, we calculated sample size needed in order for the standard error of the estimator to be at most 5% of the corresponding true parameter value for each of the 5 model parameters. The results for each of the eight scenarios in Table 1B of the main text are given in Supplementary Tables S2-S9. Each of the table contains results for each of the eight disease models given in Table 1A of the main text. As we can see from the tables, the sample size needed is the smallest for the $T + 2$ study design, whereas the P design requires the largest sample size to meet the specification. It is also seen that the homozygous genetic effect (R_2), homozygous maternal effect (S_2), and the imprinting parameter (R_{im}) are typically more difficult to estimate accurately, echoing what we saw in the simulation study (Figure 2 in the main text and Supplementary Figures S9-S17). Nevertheless, although the sample size is fairly large, they are not out of reach for consortium studies that have accumulated a large number of trio families.

Some of the results, however, are counter-intuitive and deserve further explanations. For example, the precision of the estimation of R_2 is independent of whether fathers are genotyped but dependent on the number of siblings. The reason lies in the fact that the difference between pairs (fathers not genotyped) and trios (fathers genotyped) are in the $(m, f, c) = (1, -, 1)$ genotype configurations only. In the pair data, the $(1, 1)$ configuration is ignored, whereas in the trio data, all three possible configurations are used. However, since the child's genotype is a "1", it contains no information about R_2 (informative only if a child's genotype is a "2"). Hence, the sample sizes for R_2 are the same for pair, mix, and trio. On the other hand, when additional siblings are included, they can be of the "2" genotype and thus contribute to the estimation accuracy of R_2. Thus the sample size is smaller for

families with a larger number of siblings.

As another example to show that the results are not always intuitive, we can see that the sample sizes for precise estimation of R_2 are always the same for disease models 2, 5 and 6. To explain this, we first note that only the following genotype configurations are informative for estimating R_2: $(m, f, c) = (1, 2, 2), (2, 1, 2), (1, 1, 2)$, and $(2, 2, 2)$. As such, their contributions to the likelihood do not involve the R_{im} parameter since the child has two copies of the alleles and thus not informative for R_{im}. On the other hand, although they are informative for the S_1 and S_2 parameters, for models 2, 5, and 6, there are no maternal effects. Thus the sample sizes for estimating R_2 are the same for these three disease models.

One can also see that the sample sizes for precise estimation of S_1 are the same for disease models 2, 5 and 6 when fathers are not genotyped, yet different when fathers are genotyped. For the former, this is because when father's genotype is missing, the S_1 parameter only exists in $(m, c) = (1, 2)$, and $(1, 0)$, which may also involve the R_2 parameter. However, for disease models 2, 5, and 6, the effects of R_2 are all equal to 3, thus the sample sizes should be the same for S_1. On the other hand, when father's genotype is present, the S_1 parameter is present in the following configurations: $(m, f, c) = (1, 0, 0), (1, 1, 0), (1, 0, 1), (1, 1, 1), (1, 1, 2), (1, 2, 1)$, and $(1, 2, 2)$. Their contributions to the likelihood may involve R_1 and R_{im}, which are different in models 2, 5, and 6, leading to different sample sizes for achieving the same precision. Although their contributions may also involve R_2, it is the same for all three models as pointed out earlier.

A.5: R package

An R package for calculating information and sample size is available with this Supplementary document or can be downloaded at http://www.stat.osu.edu/∼statgen/SOFTWARE/LIME.

References

CHANDA, S. (1954). A notes on the consistency and maxima of the roots of the likelihood equations.. *Biometrika* **41**, 56–61.

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.

A.5: Supplementary Tables and Figures

Supplementary Table S1: *Kolmogorov-Smirnov* distance (D) and its p-value (P) comparing empirical distribution of a parameter estimator to normal distribution.

DM Scenario[a]	Par.[b]	sample size: 200 D	200 P	1000 D	1000 P	2000 D	2000 P	10000[c] D	10000[c] P
M8S5	R_1	0.083	0.002	0.047	0.216	0.046	0.251	0.040	0.396
	R_2	0.104	$< 10^{-3}$	0.052	0.130	0.044	0.276	0.030	0.775
	S_1	0.059	0.065	0.055	0.095	0.040	0.395	0.035	0.555
	S_2	0.113	$< 10^{-3}$	0.068	0.019	0.045	0.273	0.038	0.470
	R_{im}	0.170	$< 10^{-3}$	0.089	0.001	0.065	0.028	0.041	0.364
M1S7	R_1	0.037	0.508	0.047	0.224	0.021	0.978	0.03	0.757
	R_2	0.097	$< 10^{-3}$	0.054	0.113	0.049	0.187	0.021	0.977
	S_1	0.04	0.413	0.036	0.547	0.039	0.419	0.039	0.442
	S_2	0.105	$< 10^{-3}$	0.068	0.021	0.039	0.444	0.028	0.834
	R_{im}	0.125	$< 10^{-3}$	0.054	0.111	0.053	0.119	0.033	0.64
M2S8	R_1	0.061	0.05	0.053	0.124	0.028	0.818	0.047	0.222
	R_2	0.086	0.001	0.056	0.091	0.032	0.698	0.023	0.948
	S_1	0.071	0.014	0.032	0.668	0.051	0.15	0.028	0.835
	S_2	0.103	$< 10^{-3}$	0.064	0.034	0.028	0.814	0.03	0.763
	R_{im}	0.093	$< 10^{-3}$	0.061	0.051	0.053	0.125	0.018	0.996
M3S1	R_1	0.081	0.003	0.047	0.223	0.037	0.507	0.048	0.208
	R_2	0.218	$< 10^{-3}$	0.084	0.002	0.051	0.142	0.033	0.653
	S_1	0.125	$< 10^{-3}$	0.05	0.159	0.039	0.445	0.037	0.497
	S_2	0.523	$< 10^{-3}$	0.077	0.005	0.098	$< 10^{-3}$	0.03	0.775
	R_{im}	0.28	$< 10^{-3}$	0.106	$< 10^{-3}$	0.075	0.007	0.057	0.081
M4S2	R_1	0.095	$< 10^{-3}$	0.038	0.468	0.034	0.626	0.029	0.797
	R_2	0.115	$< 10^{-3}$	0.055	0.102	0.049	0.174	0.034	0.606
	S_1	0.065	0.029	0.033	0.645	0.021	0.982	0.022	0.97
	S_2	0.189	$< 10^{-3}$	0.075	0.007	0.074	0.009	0.049	0.179
	R_{im}	0.173	$< 10^{-3}$	0.095	$< 10^{-3}$	0.079	0.004	0.028	0.842
M5S3	R_1	0.047	0.225	0.044	0.282	0.037	0.486	0.031	0.733
	R_2	0.146	$< 10^{-3}$	0.042	0.345	0.027	0.87	0.031	0.734
	S_1	0.068	0.019	0.031	0.704	0.046	0.25	0.036	0.553
	S_2	0.12	$< 10^{-3}$	0.054	0.109	0.055	0.102	0.025	0.903
	R_{im}	0.167	$< 10^{-3}$	0.064	0.035	0.04	0.391	0.035	0.576
M6S4	R_1	0.029	0.81	0.031	0.717	0.036	0.532	0.035	0.581
	R_2	0.132	$< 10^{-3}$	0.056	0.084	0.045	0.264	0.03	0.763
	S_1	0.052	0.136	0.037	0.511	0.042	0.329	0.028	0.817
	S_2	0.17	$< 10^{-3}$	0.072	0.011	0.068	0.019	0.048	0.204
	R_{im}	0.194	$< 10^{-3}$	0.063	0.038	0.066	0.026	0.036	0.537
M7S6	R_1	0.09	0.001	0.053	0.116	0.041	0.384	0.028	0.83
	R_2	0.118	$< 10^{-3}$	0.053	0.123	0.041	0.369	0.025	0.92
	S_1	0.088	0.001	0.062	0.045	0.027	0.858	0.037	0.498
	S_2	0.114	$< 10^{-3}$	0.053	0.12	0.051	0.149	0.035	0.567
	R_{im}	0.171	$< 10^{-3}$	0.067	0.023	0.045	0.252	0.036	0.526

[a]The disease models and scenarios are as defined in Table 1 of the main text: M1S7 (model 1, scenario 7), M2S8 (model 2, scenario 8), etc.

[b]R_1: relative risk of carrying one variant allele; R_2: relative risk of carry ing two variant alleles; R_{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 allele.

[c]As the sample size increases to 10,000, the empirical distribution of the estimators are indistinguishable from the normal distribution (all p-value > 0.1) under all settings considered.

Supplementary Table S2: Sample size required for a parameter estimate with standard error within 10% of true parameter value under scenario 1.

		Data Type/Study Design								
Parameter	Model	P	$P+1$	$P+2$	M	$M+1$	$M+2$	T	$T+1$	$T+2$
R_1	1	3086	2593	2236	2539	2119	1818	2006	1686	1454
	2	2351	1881	1567	1912	1520	1261	1528	1223	1019
	3	3122	2627	2268	2570	2147	1844	2029	1708	1474
	4	3044	2532	2167	2347	1919	1623	1795	1469	1243
	5	2823	2282	1915	2121	1771	1521	1615	1387	1215
	6	2220	1711	1392	1931	1319	1001	1615	1035	762
	7	2833	2175	1765	1994	1605	1343	1481	1232	1054
	8	2195	1718	1412	1811	1152	844	1466	842	591
R_2	1	36100	30337	26160	36100	30337	26160	36100	30337	26160
	2	22750	17170	13788	22750	17170	13788	22750	17170	13788
	3	21153	15501	12232	21153	15501	12232	21153	15501	12232
	4	16522	10667	7876	16522	10667	7876	16522	10667	7876
	5	22751	17170	13788	22751	17170	13788	22751	17170	13788
	6	22751	17170	13788	22751	17170	13788	22751	17170	13788
	7	18756	12975	9919	18756	12975	9919	18756	12975	9919
	8	17783	11960	9010	17783	11960	9010	17783	11960	9010
S_1	1	5731	4816	4153	4093	3401	2909	2866	2408	2077
	2	5933	4946	4240	3756	3044	2559	2515	2047	1726
	3	5404	4442	3771	3939	3224	2729	2797	2325	1990
	4	4098	3161	2573	2927	2260	1841	2115	1661	1367
	5	5933	4946	4240	3499	3043	2693	2308	2061	1862
	6	5933	4946	4240	4239	2659	1937	2967	1713	1204
	7	4680	3740	3114	2762	2321	2002	1866	1614	1422
	8	4414	3478	2869	3095	1743	1213	2207	1119	750
S_2	1	9757	8199	7071	9757	8199	7071	9757	8199	7071
	2	7281	5784	4797	7281	5784	4797	7281	5784	4797
	3	9202	7563	6420	9202	7563	6420	9202	7563	6420
	4	6977	5382	4381	6977	5382	4381	6977	5382	4381
	5	6149	4641	3727	6149	4641	3727	6149	4641	3727
	6	10102	8421	7219	10102	8421	7219	10102	8421	7219
	7	5069	3507	2681	5069	3507	2681	5069	3507	2681
	8	7516	5922	4885	7516	5922	4885	7516	5922	4885
R_{im}	1	10841	9110	7856	6449	5750	5187	4012	3711	3452
	2	8259	6607	5506	4784	4139	3647	3056	2762	2519
	3	10967	9230	7967	6531	5828	5261	4058	3756	3496
	4	8269	6617	5516	4790	4145	3653	3060	2765	2523
	5	6832	5157	4141	3883	3239	2778	2528	2220	1979
	6	12088	10287	8954	7261	6523	5921	4473	4162	3892
	7	5633	3897	2979	3148	2482	2049	2084	1752	1510
	8	8909	7249	6111	5200	4546	4038	3297	3003	2758

Supplementary Table S3: Sample size required for parameter estimation with standard error within 10% of true parameter value under scenario 2.

		Data Type/Study Design								
Parameter	Model	P	$P+1$	$P+2$	M	$M+1$	$M+2$	T	$T+1$	$T+2$
R_1	1	4012	3371	2907	2866	2381	2037	2006	1686	1454
	2	3056	2370	1936	2142	1642	1332	1528	1185	968
	3	4058	3418	2952	2901	2410	2061	2029	1703	1468
	4	4338	3667	3175	2717	2151	1780	1806	1418	1167
	5	4153	3388	2860	2449	2081	1809	1615	1410	1252
	6	2643	1856	1430	2107	1286	926	1615	940	663
	7	4640	3501	2811	2350	1876	1561	1488	1224	1039
	8	2787	2068	1644	2021	1062	720	1474	691	451
R_2	1	36101	30337	26160	36101	30337	26160	36100	30337	26160
	2	22751	15794	12096	22751	15794	12096	22751	15794	12096
	3	21153	14451	10974	21153	14451	10974	21153	14451	10974
	4	16710	8569	5762	16710	8569	5762	16710	8569	5762
	5	22751	15794	12096	22751	15794	12096	22751	15794	12096
	6	22751	15794	12096	22751	15794	12096	22751	15794	12096
	7	18904	10612	7376	18904	10612	7376	18904	10612	7376
	8	17979	9375	6341	17979	9375	6341	17979	9375	6341
S_1	1	4012	3371	2907	2866	2381	2037	2006	1686	1454
	2	4153	3356	2815	2629	2062	1696	1761	1389	1147
	3	3783	3068	2580	2758	2231	1873	1958	1612	1370
	4	2900	2039	1573	2074	1465	1133	1497	1084	849
	5	4153	3289	2723	2449	2044	1753	1615	1393	1225
	6	4153	3425	2913	2968	1885	1381	2077	1224	867
	7	3305	2316	1782	1951	1472	1182	1317	1038	856
	8	3127	2261	1770	2194	1110	743	1564	711	460
S_2	1	36101	30337	26160	36101	30337	26160	36100	30337	26160
	2	26938	19905	15784	26938	19905	15784	26938	19905	15784
	3	34045	27194	22638	34045	27194	22638	34045	27194	22638
	4	26099	17803	13509	26099	17803	13509	26099	17803	13509
	5	22751	14622	10773	22751	14622	10773	22751	14622	10773
	6	37375	30817	26216	37375	30817	26216	37375	30817	26216
	7	18904	9372	6230	18904	9372	6230	18904	9372	6230
	8	28136	20342	15930	28136	20342	15930	28136	20342	15930
R_{im}	1	40112	33707	29067	8720	8228	7788	4012	3872	3742
	2	30556	22908	18322	6314	5742	5265	3056	2875	2714
	3	40576	33948	29182	8839	8299	7821	4058	3901	3756
	4	30930	23110	18447	6406	5778	5262	3093	2888	2708
	5	25278	16247	11970	5041	4386	3882	2528	2311	2128
	6	44723	38287	33471	9907	9345	8843	4473	4311	4160
	7	21005	10413	6923	4065	3220	2665	2101	1796	1569
	8	33355	25979	21275	7009	6376	5848	3336	3129	2947

Supplementary Table S4: Sample size required for parameter estimation with standard error within 10% of true parameter value under scenario 3.

		Data Type/Study Design								
Parameter	Model	P	$P+1$	$P+2$	M	$M+1$	$M+2$	T	$T+1$	$T+2$
R_1	1	2471	1636	1223	2033	1327	985	1606	1064	795
	2	1667	944	658	1353	759	527	1084	617	431
	3	2508	1665	1246	2065	1347	1000	1631	1078	806
	4	2372	1471	1066	1785	1043	737	1344	780	549
	5	2025	1078	734	1452	855	606	1081	688	505
	6	1411	701	466	1257	540	344	1081	431	269
	7	1631	600	367	1046	467	300	749	377	252
	8	1467	778	530	1196	433	265	966	297	175
R_2	1	28900	19140	14307	28900	19140	14307	28900	19140	14307
	2	13751	6454	4216	13751	6454	4216	13751	6454	4216
	3	11624	5255	3395	11624	5255	3395	11624	5255	3395
	4	4537	1309	765	4537	1309	765	4537	1309	765
	5	13750	6377	4152	13750	6377	4152	13750	6377	4152
	6	13750	6532	4283	13750	6532	4283	13750	6532	4283
	7	8149	2688	1610	8149	2688	1610	8149	2688	1610
	8	6624	2031	1200	6624	2031	1200	6624	2031	1200
S_1	1	4588	3038	2271	3277	2119	1566	2294	1519	1136
	2	4623	2807	2015	2813	1627	1144	1854	1090	771
	3	4042	2455	1763	3003	1821	1307	2164	1362	994
	4	2331	1012	646	1761	789	508	1331	629	412
	5	4623	2720	1927	2433	1641	1238	1545	1120	878
	6	4623	2900	2112	3294	1297	808	2312	808	489
	7	3184	1480	965	1523	869	608	966	601	437
	8	2830	1317	858	1968	561	327	1415	351	201
S_2	1	7811	5173	3867	7811	5173	3867	7811	5173	3867
	2	5059	2677	1820	5059	2677	1820	5059	2677	1820
	3	6882	4043	2862	6882	4043	2862	6882	4043	2862
	4	3969	1590	994	3969	1590	994	3969	1590	994
	5	3717	1592	1013	3717	1592	1013	3717	1592	1013
	6	7871	4937	3596	7871	4937	3596	7871	4937	3596
	7	2203	648	380	2203	648	380	2203	648	380
	8	4818	2241	1461	4818	2241	1461	4818	2241	1461
R_{im}	1	8679	5748	4297	5163	3892	3123	3212	2638	2238
	2	5856	3186	2188	3379	2222	1656	2167	1597	1264
	3	8812	5797	4319	5248	3924	3134	3261	2658	2244
	4	5868	3145	2148	3386	2190	1619	2171	1574	1234
	5	4130	1769	1126	2328	1295	897	1528	984	726
	6	9985	6853	5217	6004	4609	3740	3695	3070	2626
	7	2448	720	423	1351	563	356	906	455	304
	8	6591	3788	2658	3837	2596	1962	2439	1828	1462

Supplementary Table S5: Sample size required for parameter estimation with standard error within 10% of true parameter value under scenario 4.

		Data Type/Study Design								
Parameter	Model	P	$P+1$	$P+2$	M	$M+1$	$M+2$	T	$T+1$	$T+2$
R_1	1	3212	2127	1590	2294	1484	1096	1606	1064	795
	2	2167	1246	875	1515	854	594	1084	623	438
	3	3261	2168	1624	2331	1507	1114	1631	1078	805
	4	3519	2345	1758	2090	1230	872	1358	792	559
	5	3236	1984	1430	1703	1173	895	1081	794	627
	6	1623	791	523	1350	536	334	1081	395	242
	7	3304	1599	1055	1285	776	556	766	499	370
	8	1842	983	671	1320	416	247	966	260	150
R_2	1	28900	19140	14307	28900	19140	14307	28900	19140	14307
	2	13750	6612	4352	13750	6612	4352	13750	6612	4352
	3	11624	5446	3556	11624	5446	3556	11624	5446	3556
	4	4857	1517	899	4857	1517	899	4857	1517	899
	5	13750	6612	4352	13750	6612	4352	13750	6612	4352
	6	13750	6612	4352	13750	6612	4352	13750	6612	4352
	7	8376	2959	1797	8376	2959	1797	8376	2959	1797
	8	6937	2218	1321	6937	2218	1321	6937	2218	1321
S_1	1	3212	2127	1590	2294	1484	1096	1606	1064	795
	2	3236	1967	1413	1969	1140	802	1298	763	541
	3	2829	1721	1237	2102	1277	917	1515	955	697
	4	1688	753	485	1271	583	379	957	462	305
	5	3236	1904	1349	1703	1145	862	1081	781	611
	6	3236	2035	1484	2306	913	570	1618	569	345
	7	2266	1067	698	1091	625	438	692	432	314
	8	2032	970	637	1415	414	242	1016	259	149
S_2	1	28900	19140	14307	28900	19140	14307	28900	19140	14307
	2	18718	9906	6735	18718	9906	6735	18718	9906	6735
	3	25461	15008	10640	25461	15008	10640	25461	15008	10640
	4	15186	6302	3976	15186	6302	3976	15186	6302	3976
	5	13751	5890	3748	13751	5890	3748	13751	5890	3748
	6	29121	18312	13356	29121	18312	13356	29121	18312	13356
	7	8376	2508	1475	8376	2508	1475	8376	2508	1475
	8	18280	8723	5729	18280	8723	5729	18280	8723	5729
R_{im}	1	32112	21266	15897	6981	6014	5283	3212	2928	2690
	2	21667	11779	8088	4443	3559	2968	2167	1871	1646
	3	32602	21417	15947	7104	6056	5278	3261	2947	2688
	4	22095	11945	8185	4543	3580	2953	2210	1878	1633
	5	15278	6545	4164	3001	2215	1755	1528	1245	1050
	6	36945	25327	19268	8201	7056	6191	3695	3358	3077
	7	9307	2787	1639	1761	1105	805	931	668	521
	8	24821	14378	10120	5191	4155	3464	2483	2130	1865

Supplementary Table S6: Sample size required for parameter estimation with standard error within 10% of true parameter value under scenario 5.

		Data Type/Study Design								
Parameter	Model	P	$P+1$	$P+2$	M	$M+1$	$M+2$	T	$T+1$	$T+2$
R_1	1	1323	1112	959	1089	908	779	860	723	623
	2	1168	951	801	956	773	649	759	619	522
	3	1460	1247	1088	1206	1020	884	949	807	702
	4	1589	1349	1172	1244	1033	884	943	784	670
	5	1325	1047	865	1042	841	705	806	667	569
	6	1164	908	745	988	726	574	806	579	452
	7	1453	1075	854	1088	833	675	817	648	537
	8	1246	1032	881	1007	736	579	794	547	418
R_2	1	4012	3371	2907	4012	3371	2907	4012	3371	2907
	2	2887	2162	1728	2887	2162	1728	2887	2162	1728
	3	2509	1741	1333	2509	1741	1333	2509	1741	1333
	4	2227	1362	981	2227	1362	981	2227	1362	981
	5	2887	2153	1716	2887	2153	1716	2887	2153	1716
	6	2887	2171	1739	2887	2171	1739	2887	2171	1739
	7	2643	1844	1417	2643	1844	1417	2643	1844	1417
	8	2488	1660	1246	2488	1660	1246	2488	1660	1246
S_1	1	2456	2064	1780	1755	1458	1247	1228	1032	890
	2	2617	2132	1798	1748	1407	1177	1186	967	817
	3	2205	1738	1434	1673	1327	1099	1216	988	832
	4	1916	1402	1106	1435	1064	845	1049	801	648
	5	2617	2127	1791	1702	1419	1217	1151	986	862
	6	2617	2137	1805	1868	1344	1050	1309	913	701
	7	2348	1832	1501	1523	1225	1025	1040	862	736
	8	2182	1653	1330	1541	1004	745	1091	684	498
S_2	1	2360	1983	1710	2360	1983	1710	2360	1983	1710
	2	1951	1520	1245	1951	1520	1245	1951	1520	1245
	3	2119	1613	1303	2119	1613	1303	2119	1613	1303
	4	1841	1283	984	1841	1283	984	1841	1283	984
	5	1698	1231	965	1698	1231	965	1698	1231	965
	6	2514	2053	1734	2514	2053	1734	2514	2053	1734
	7	1555	1046	788	1555	1046	788	1555	1046	788
	8	2097	1588	1278	2097	1588	1278	2097	1588	1278
R_{im}	1	3371	2833	2443	2435	2109	1859	1720	1533	1382
	2	2977	2389	1995	2133	1780	1527	1518	1312	1156
	3	3720	3146	2725	2704	2348	2075	1898	1695	1532
	4	3182	2570	2156	2290	1915	1646	1623	1404	1237
	5	2426	1758	1379	1714	1323	1077	1238	1003	844
	6	4523	3964	3528	3323	2966	2679	2307	2107	1939
	7	2221	1494	1126	1560	1134	890	1133	873	710
	8	3677	3101	2680	2670	2307	2031	1875	1665	1497

Supplementary Table S7: Sample size required for parameter estimation with standard error within 10% of true parameter value under scenario 6.

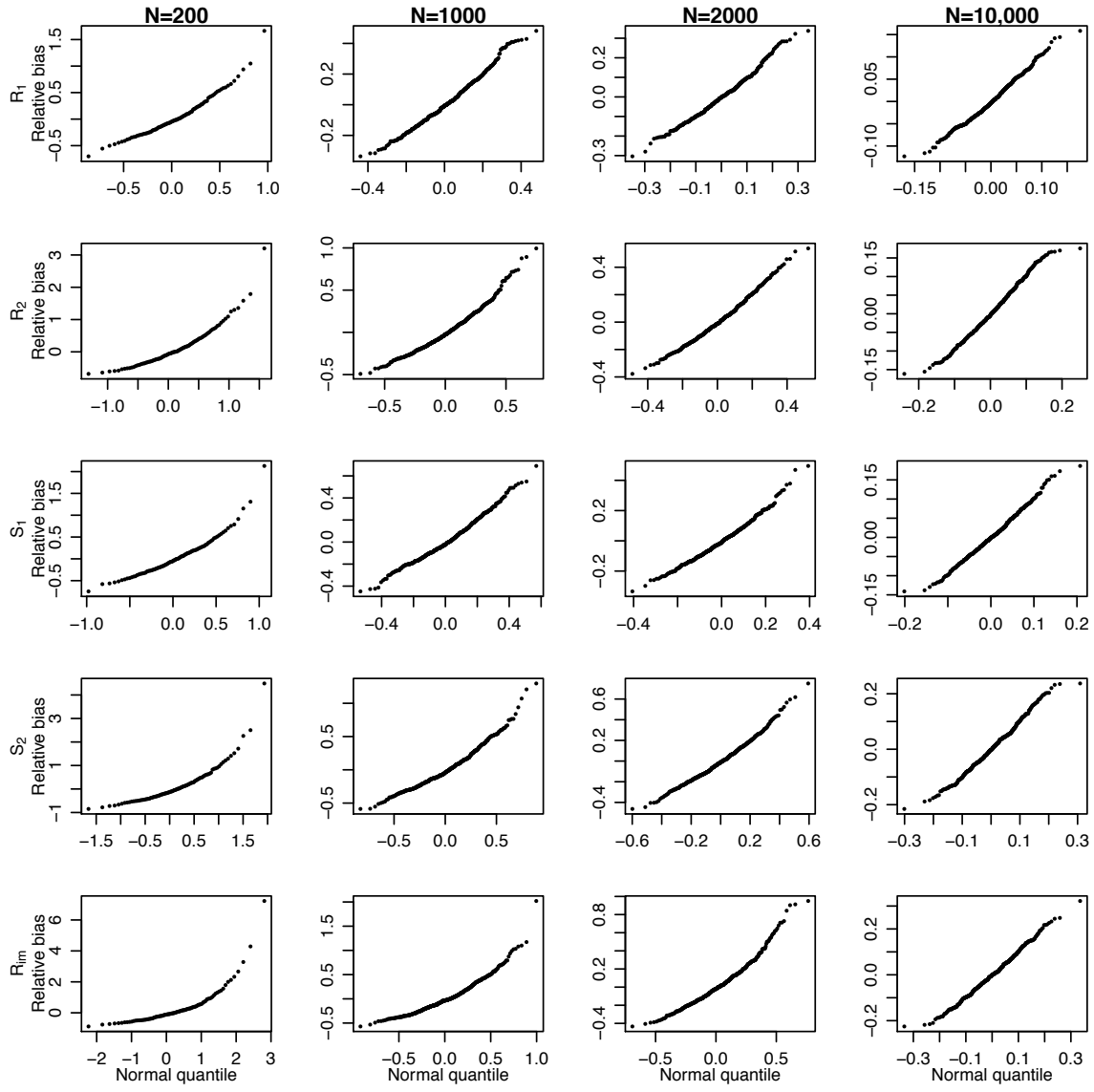
		Data Type/Study Design								
Parameter	Model	P	$P+1$	$P+2$	M	$M+1$	$M+2$	T	$T+1$	$T+2$
R_1	1	1720	1445	1246	1228	1021	873	860	723	623
	2	1518	1239	1047	1076	868	728	759	620	524
	3	1898	1622	1416	1364	1148	991	949	806	701
	4	2203	1896	1664	1433	1188	1015	945	786	673
	5	1832	1477	1238	1191	986	842	806	685	596
	6	1437	1114	910	1094	777	603	806	561	430
	7	2130	1633	1325	1259	1006	838	816	675	575
	8	1625	1354	1161	1136	792	608	797	527	394
R_2	1	4012	3371	2907	4012	3371	2907	4012	3371	2907
	2	2887	2180	1751	2887	2180	1751	2887	2180	1751
	3	2509	1771	1369	2509	1771	1369	2509	1771	1369
	4	2261	1442	1059	2261	1442	1059	2261	1442	1059
	5	2887	2180	1751	2887	2180	1751	2887	2180	1751
	6	2887	2180	1751	2887	2180	1751	2887	2180	1751
	7	2672	1911	1488	2672	1911	1488	2672	1911	1488
	8	2536	1735	1318	2536	1735	1318	2536	1735	1318
S_1	1	1720	1445	1246	1228	1021	873	860	723	623
	2	1832	1493	1260	1224	985	825	831	678	572
	3	1544	1218	1006	1171	930	771	851	692	583
	4	1365	1012	804	1024	768	614	748	577	470
	5	1832	1489	1254	1191	992	849	806	688	600
	6	1832	1498	1267	1308	947	742	916	644	496
	7	1666	1307	1075	1081	872	730	737	612	522
	8	1563	1201	975	1106	737	553	782	502	370
S_2	1	4012	3371	2907	4012	3371	2907	4012	3371	2907
	2	3316	2584	2117	3316	2584	2117	3316	2584	2117
	3	3601	2748	2222	3601	2748	2222	3601	2748	2222
	4	3185	2255	1746	3185	2255	1746	3185	2255	1746
	5	2887	2092	1640	2887	2092	1640	2887	2092	1640
	6	4274	3494	2955	4274	3494	2955	4274	3494	2955
	7	2672	1816	1375	2672	1816	1375	2672	1816	1375
	8	3647	2801	2274	3647	2801	2274	3647	2801	2274
R_{im}	1	5731	4816	4153	2964	2617	2343	1720	1568	1440
	2	5060	4058	3387	2581	2209	1931	1518	1351	1217
	3	6324	5338	4618	3306	2918	2611	1898	1729	1589
	4	5514	4474	3764	2840	2438	2135	1655	1474	1329
	5	4124	2988	2343	2053	1654	1384	1238	1051	914
	6	7689	6731	5985	4096	3695	3366	2307	2137	1991
	7	3816	2594	1965	1882	1453	1183	1145	940	797
	8	6416	5433	4712	3359	2955	2638	1925	1746	1598

Supplementary Table S8: Sample size required for parameter estimation with standard error within 10% of true parameter value under scenario 7.

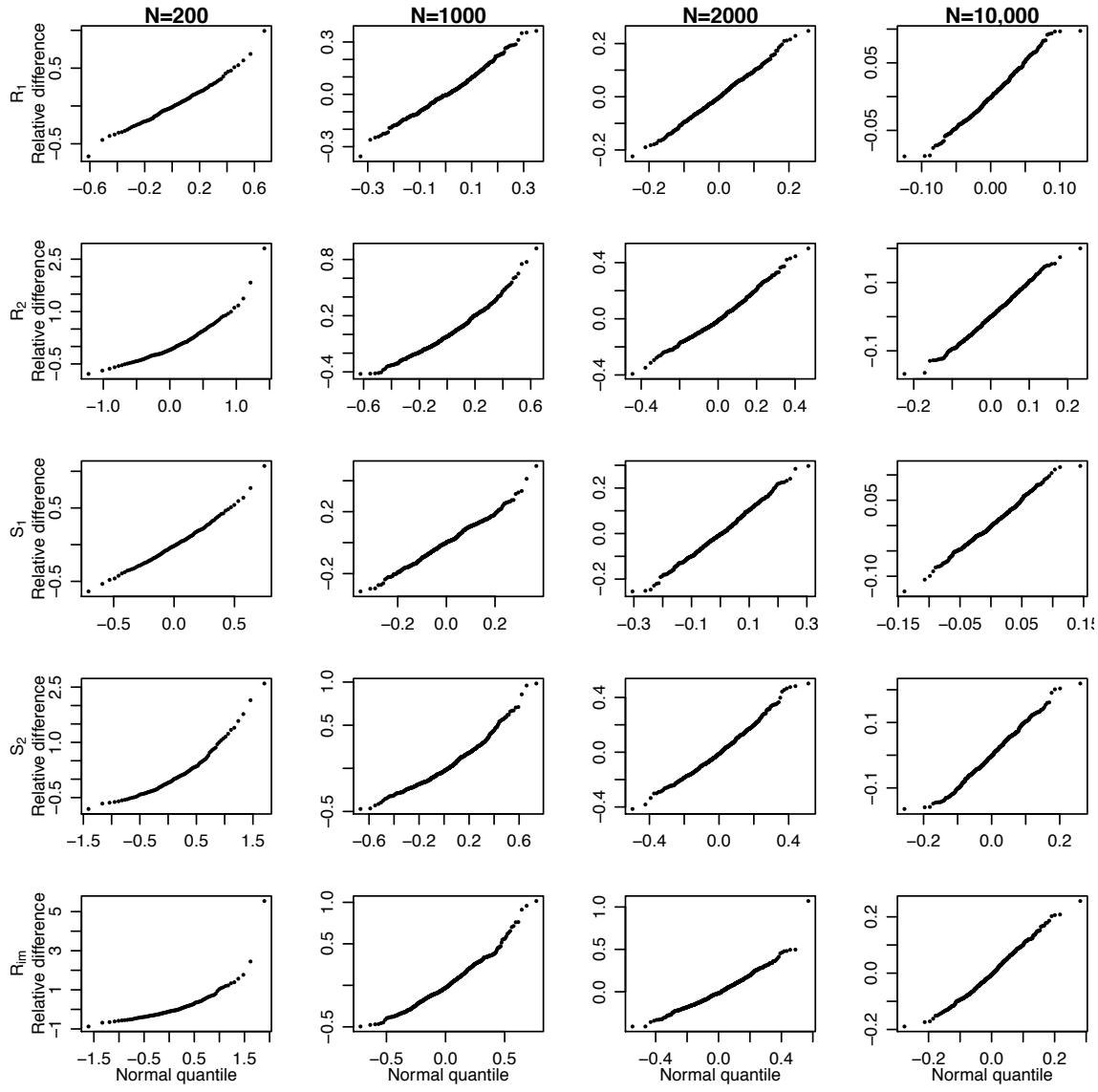
		Data Type/Study Design								
Parameter	Model	P	$P+1$	$P+2$	M	$M+1$	$M+2$	T	$T+1$	$T+2$
R_1	1	1059	702	525	872	569	422	689	456	341
	2	893	558	406	731	451	326	581	364	265
	3	1203	821	623	994	665	499	782	530	400
	4	1314	883	665	1016	655	483	763	494	365
	5	998	583	412	772	474	342	593	384	284
	6	839	483	339	718	373	252	593	295	196
	7	1027	530	358	754	419	290	562	335	239
	8	971	625	461	766	388	260	595	274	178
R_2	1	3212	2127	1590	3212	2127	1590	3212	2127	1590
	2	1970	1069	734	1970	1069	734	1970	1069	734
	3	1504	722	475	1504	722	475	1504	722	475
	4	1117	445	278	1117	445	278	1117	445	278
	5	1970	1062	727	1970	1062	727	1970	1062	727
	6	1970	1076	740	1970	1076	740	1970	1076	740
	7	1674	812	536	1674	812	536	1674	812	536
	8	1476	668	432	1476	668	432	1476	668	432
S_1	1	1966	1302	974	1405	909	671	983	651	487
	2	2018	1248	903	1343	814	584	911	568	413
	3	1564	904	636	1228	718	508	920	563	406
	4	1215	609	406	956	491	330	729	394	270
	5	2018	1245	901	1268	834	621	847	590	453
	6	2018	1251	906	1436	707	469	1009	471	307
	7	1726	987	691	1072	661	478	722	474	352
	8	1538	826	565	1080	457	290	769	307	192
S_2	1	1889	1251	936	1889	1251	936	1889	1251	936
	2	1433	827	582	1433	827	582	1433	827	582
	3	1503	816	560	1503	816	560	1503	816	560
	4	1167	536	348	1167	536	348	1167	536	348
	5	1159	597	402	1159	597	402	1159	597	402
	6	1939	1202	871	1939	1202	871	1939	1202	871
	7	985	452	293	985	452	293	985	452	293
	8	1478	794	543	1478	794	543	1478	794	543
R_{im}	1	2699	1788	1336	1950	1377	1064	1377	1037	832
	2	2276	1388	999	1629	1075	802	1161	825	639
	3	3065	2053	1544	2229	1583	1227	1563	1183	951
	4	2498	1541	1114	1797	1190	890	1274	906	703
	5	1656	852	574	1166	672	473	845	534	390
	6	3893	2791	2176	2864	2144	1714	1986	1571	1299
	7	1407	645	419	983	515	349	718	416	293
	8	3020	2013	1509	2195	1543	1189	1540	1151	919

Supplementary Table S9: Sample size required for parameter estimation with standard error within 10% of true parameter value under scenario 8.

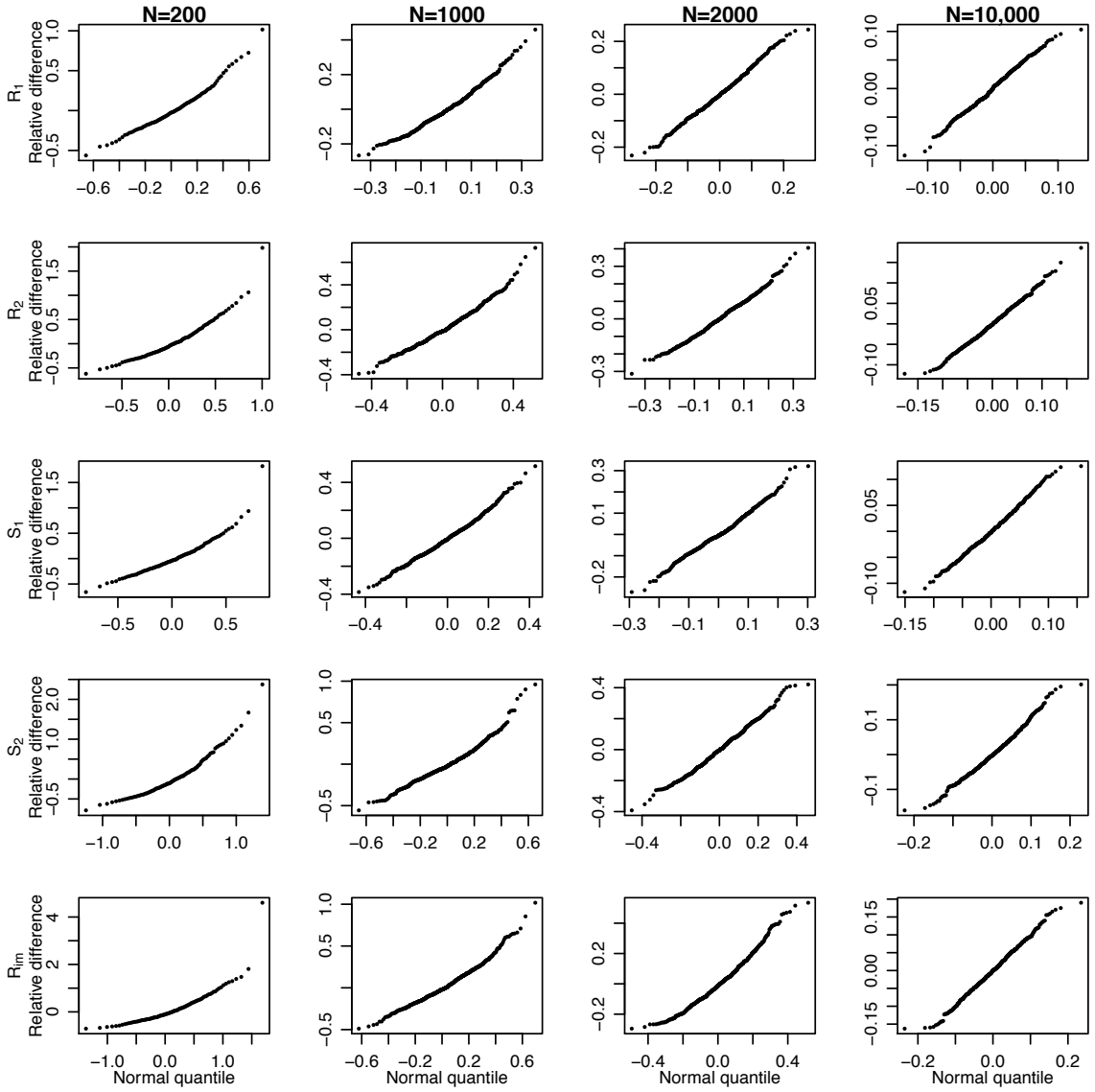
		Study Design								
Parameter	Model	P	$P+1$	$P+2$	M	$M+1$	$M+2$	T	$T+1$	$T+2$
R_1	1	1377	912	682	983	636	470	689	456	341
	2	1161	730	532	822	505	365	581	365	266
	3	1563	1069	812	1125	744	556	782	528	399
	4	1851	1276	974	1178	755	555	767	497	368
	5	1413	854	612	887	573	423	593	406	309
	6	1021	582	407	789	388	257	593	280	183
	7	1569	854	587	884	533	382	567	368	272
	8	1269	827	613	858	401	262	594	256	163
R_2	1	3212	2127	1590	3212	2127	1590	3212	2127	1590
	2	1970	1084	747	1970	1084	747	1970	1084	747
	3	1504	742	493	1504	742	493	1504	742	493
	4	1167	494	313	1167	494	313	1167	494	313
	5	1970	1084	747	1970	1084	747	1970	1084	747
	6	1970	1084	747	1970	1084	747	1970	1084	747
	7	1710	862	576	1710	862	576	1710	862	576
	8	1538	723	473	1538	723	473	1538	723	473
S_1	1	1377	912	682	983	636	470	689	456	341
	2	1413	875	634	940	571	410	638	398	290
	3	1095	634	447	860	504	356	644	395	285
	4	881	453	305	693	364	247	527	291	201
	5	1413	872	631	887	581	432	593	410	314
	6	1413	878	637	1005	500	333	707	334	219
	7	1232	712	501	767	474	343	516	339	252
	8	1118	617	426	786	344	221	559	231	146
S_2	1	3212	2127	1590	3212	2127	1590	3212	2127	1590
	2	2435	1407	990	2435	1407	990	2435	1407	990
	3	2554	1392	957	2554	1392	957	2554	1392	957
	4	2055	974	638	2055	974	638	2055	974	638
	5	1970	1014	683	1970	1014	683	1970	1014	683
	6	3296	2049	1486	3296	2049	1486	3296	2049	1486
	7	1710	797	519	1710	797	519	1710	797	519
	7	1710	797	519	1710	797	519	1710	797	519
	8	2607	1438	993	2607	1438	993	2607	1438	993
R_{im}	1	4588	3038	2271	2373	1750	1386	1377	1093	906
	2	3870	2357	1694	1970	1371	1051	1161	878	706
	3	5210	3478	2611	2727	2008	1589	1563	1240	1027
	4	4358	2709	1965	2244	1568	1205	1308	991	797
	5	2815	1448	975	1393	871	634	845	585	448
	6	6618	4734	3685	3534	2716	2205	1986	1629	1380
	7	2442	1138	742	1195	696	491	733	477	354
	8	5306	3564	2683	2781	2035	1605	1592	1251	1030



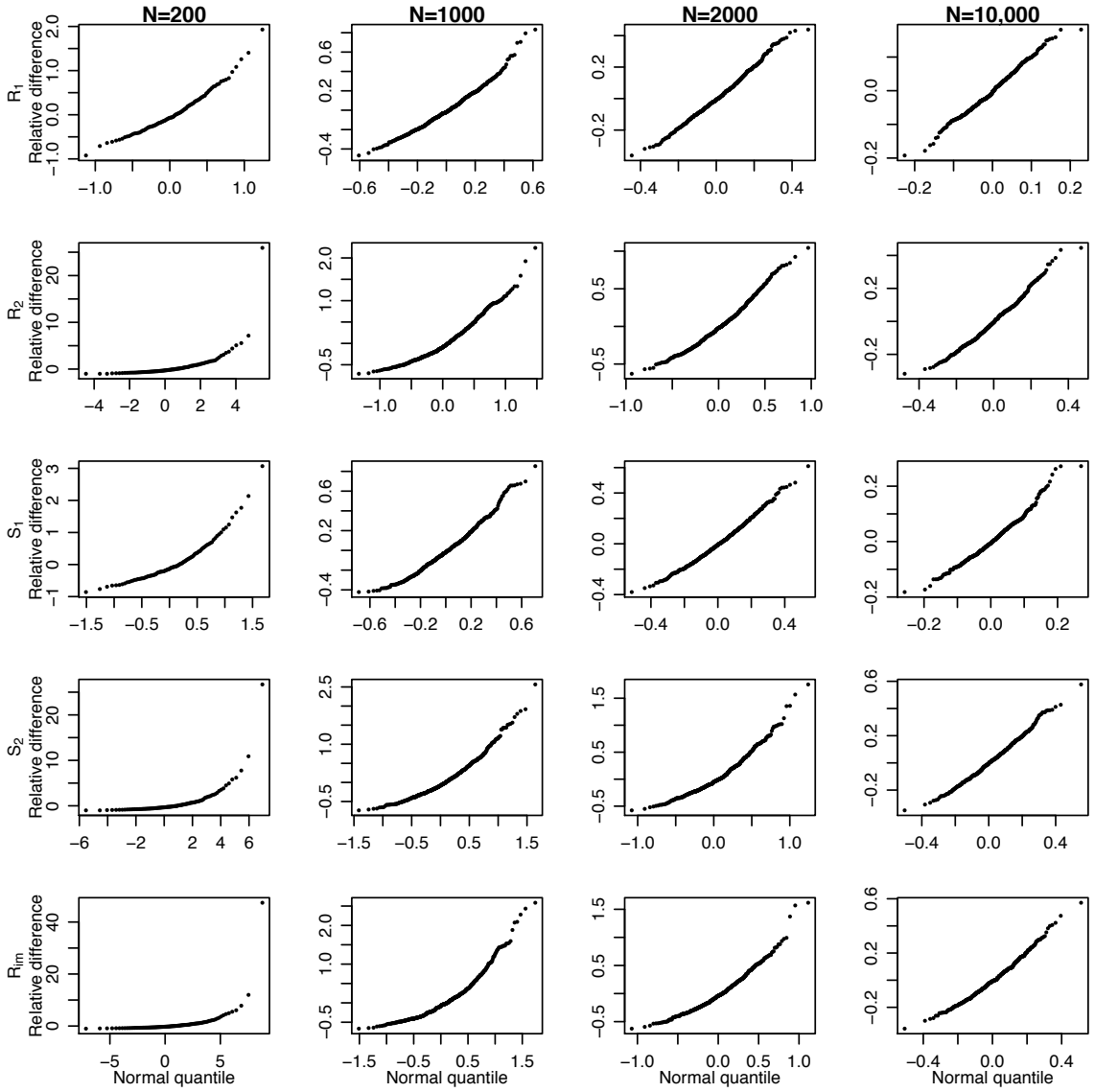
Supplementary Figure S1: QQ-norm plots for relative distance of a parameter and its LIME estimator with varying number of nuclear families: N = 200, 1000, 2000 and 10000, for disease model 8 and scenario 5.



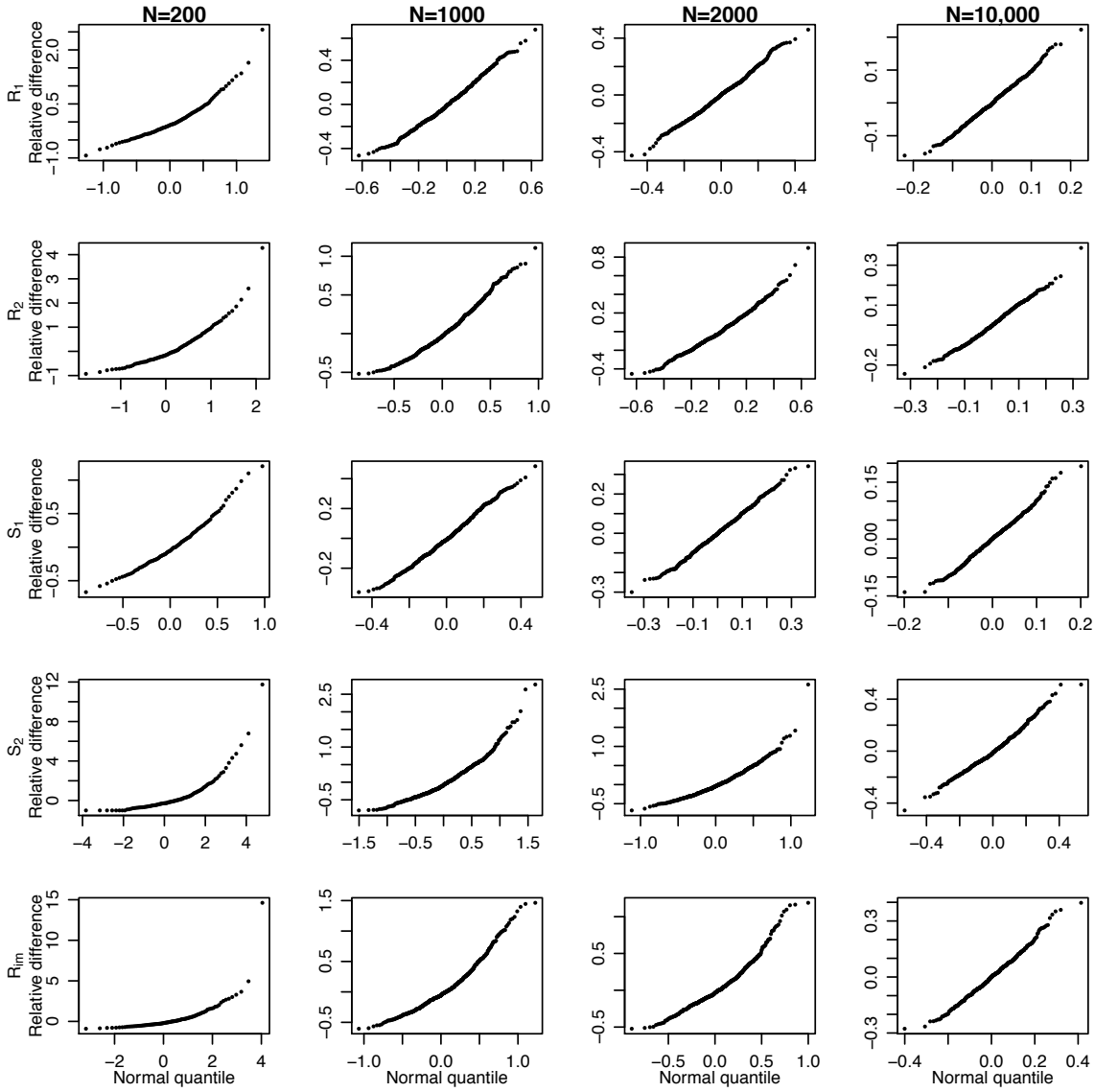
Supplementary Figure S2: QQ-norm plots for relative distance of a parameter and its LIME estimator with varying number of nuclear families: N = 200, 1000, 2000 and 10000, for disease model 1 and scenario 7.



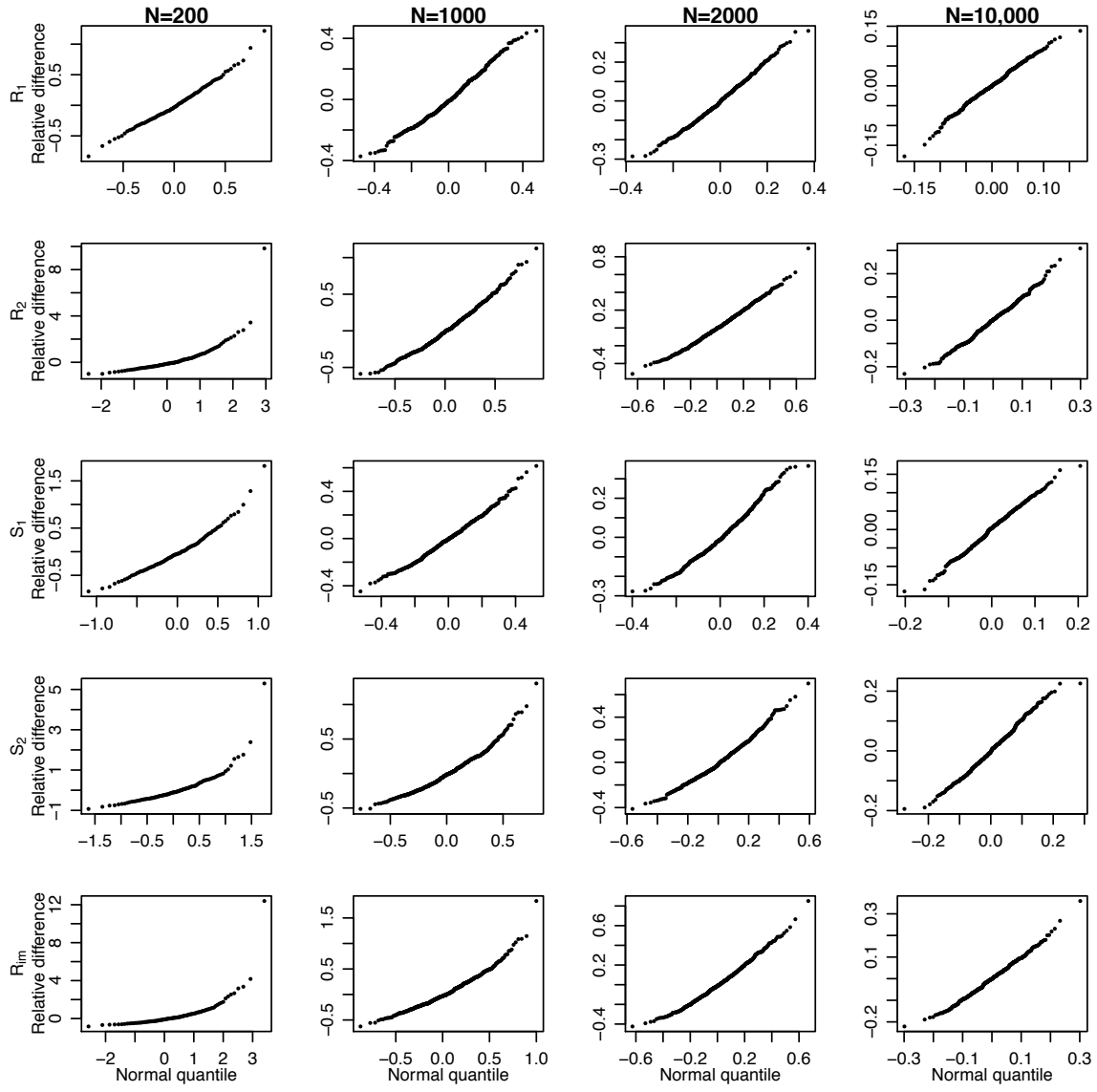
Supplementary Figure S3: QQ-norm plots for relative distance of a parameter and its LIME estimator with varying number of nuclear families: N = 200, 1000, 2000 and 10000, for disease model 2 and scenario 8.



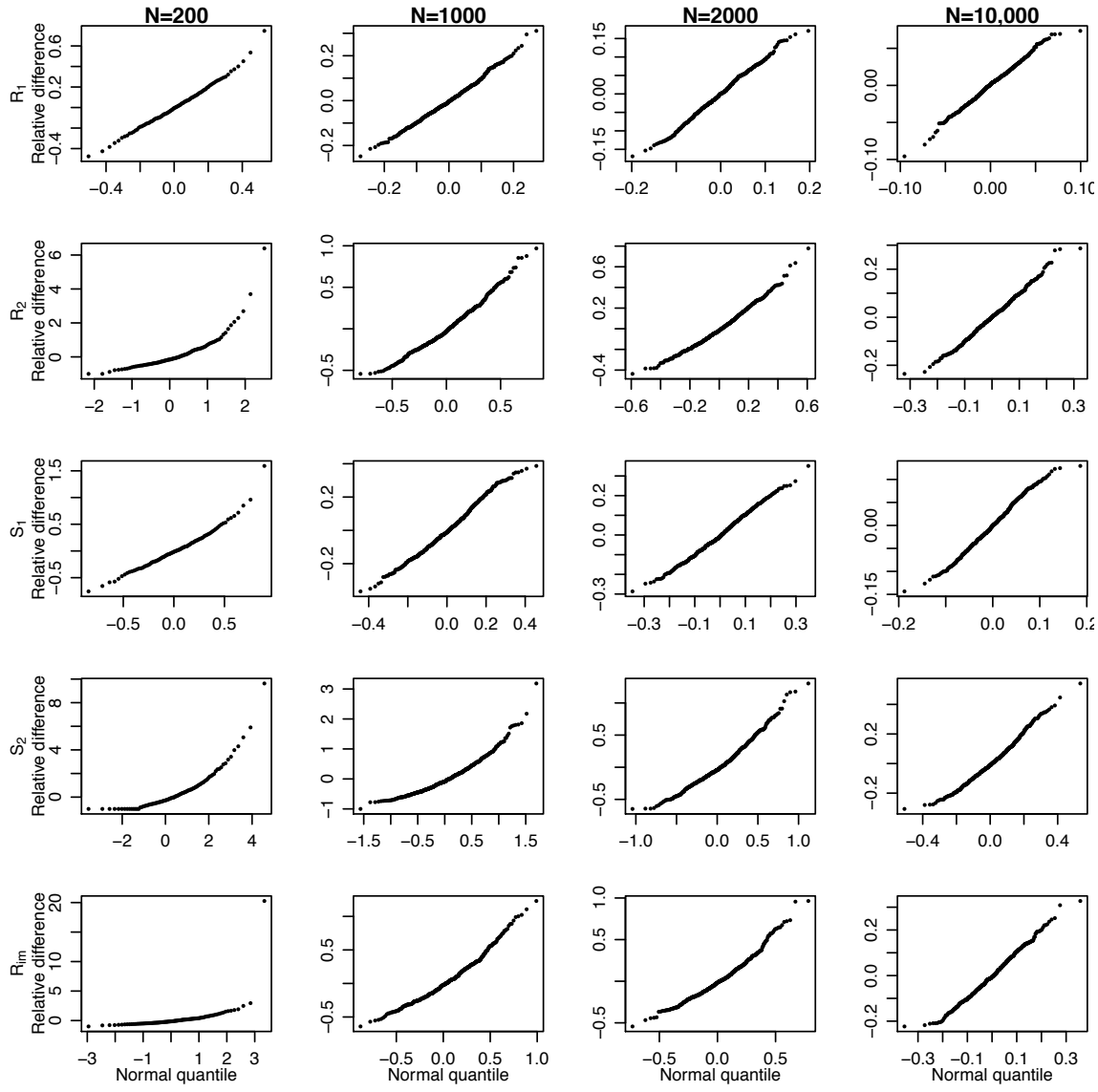
Supplementary Figure S4: QQ-norm plots for relative distance of a parameter and its LIME estimator with varying number of nuclear families: N = 200, 1000, 2000 and 10000, for disease model 3 and scenario 1.



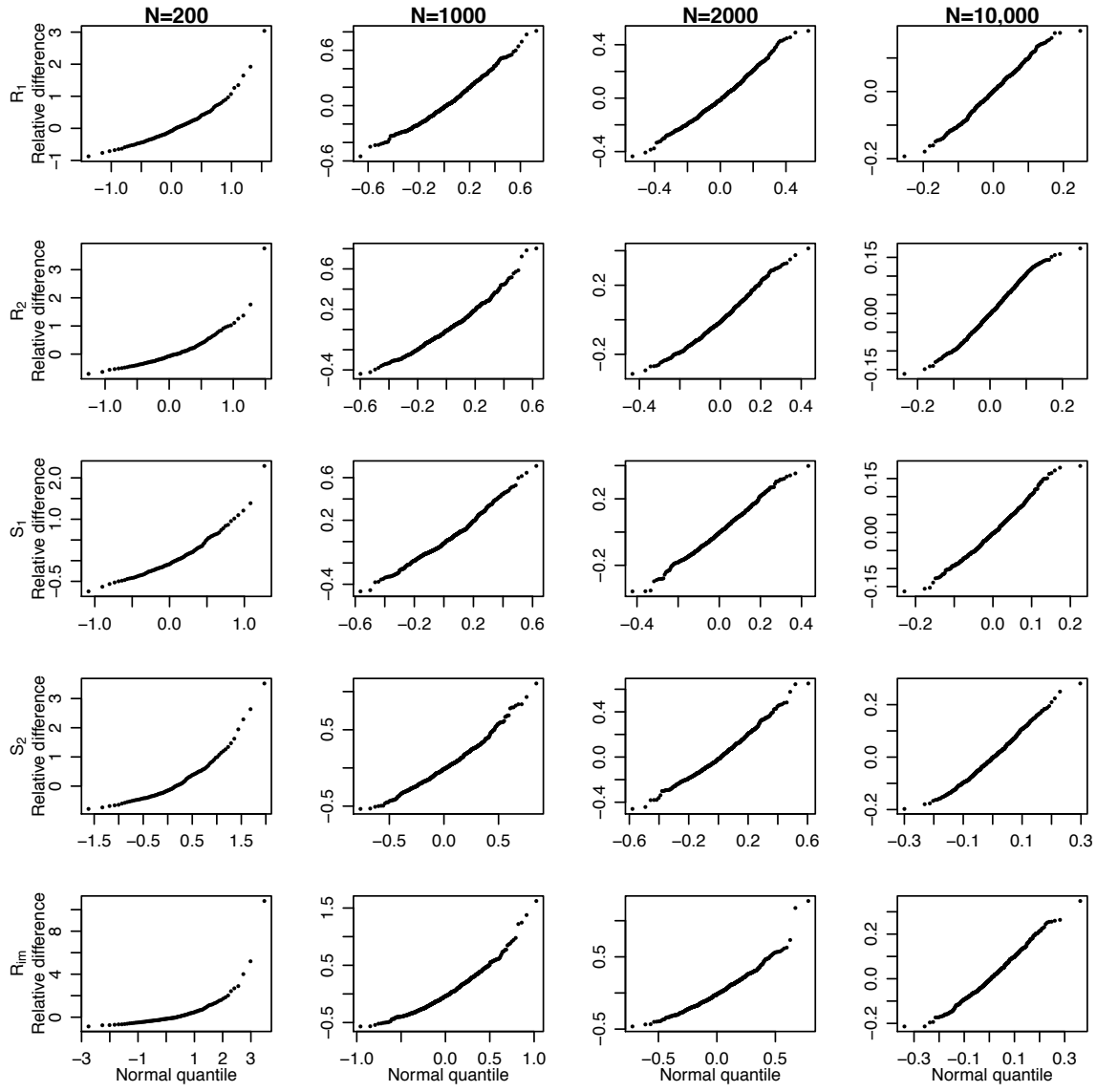
Supplementary Figure S5: QQ-norm plots for relative distance of a parameter and its LIME estimator with varying number of nuclear families: N = 200, 1000, 2000 and 10000, for disease model 4 and scenario 2.



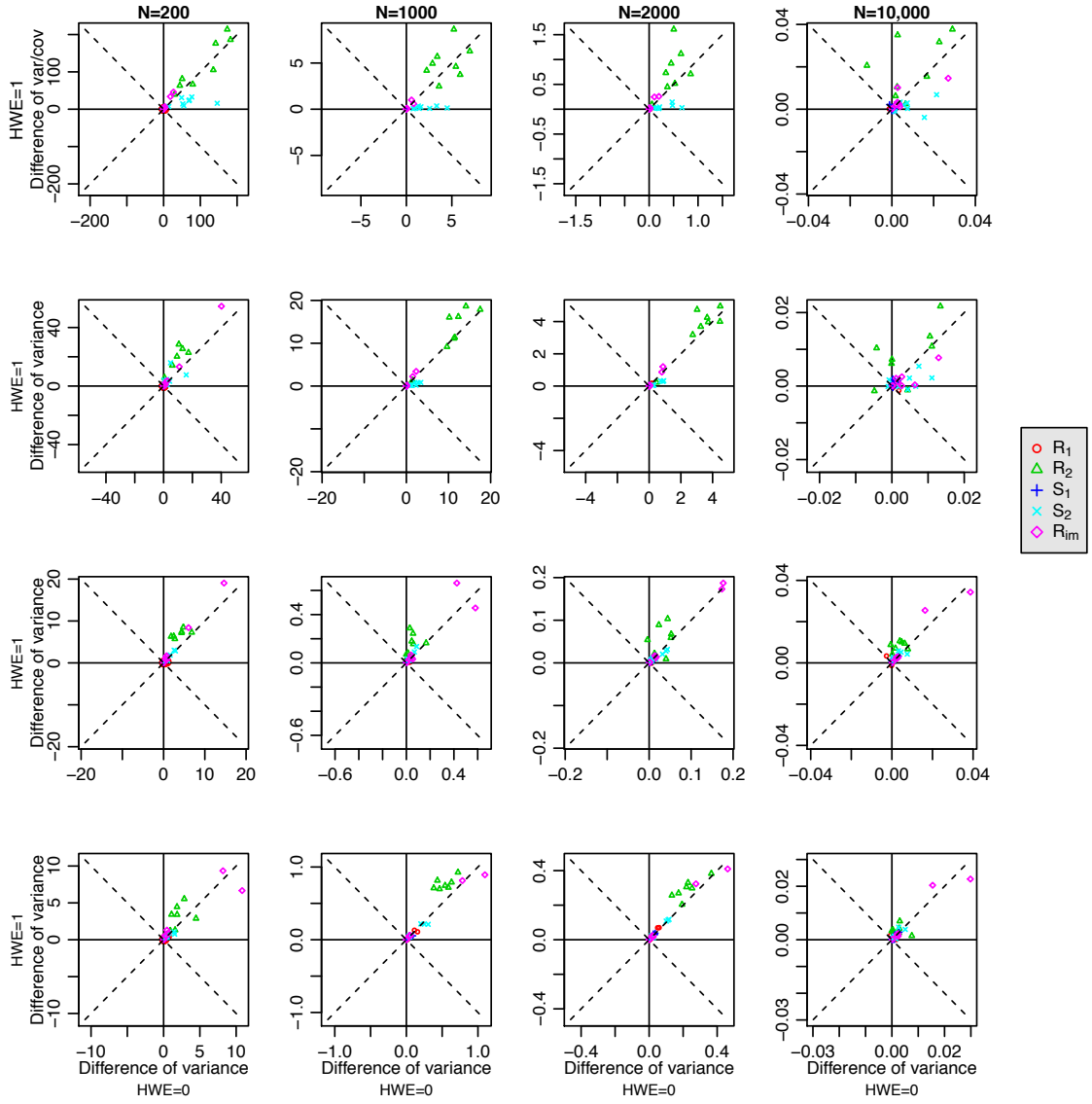
Supplementary Figure S6: QQ-norm plots for relative distance of a parameter and its LIME estimator with varying number of nuclear families: N = 200, 1000, 2000 and 10000, for disease model 5 and scenario 3.



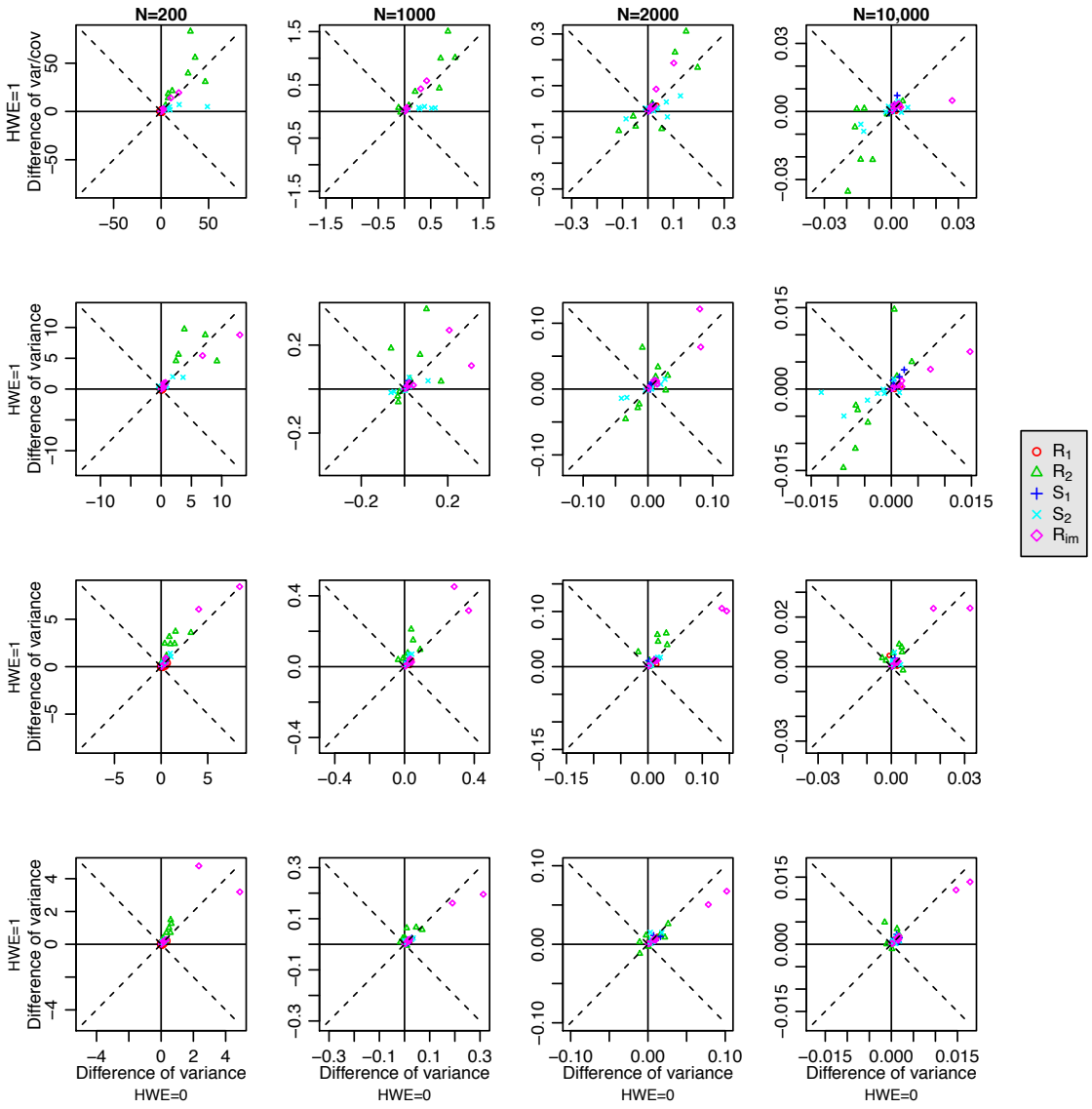
Supplementary Figure S7: QQ-norm plots for relative distance of a parameter and its LIME estimator with varying number of nuclear families: N = 200, 1000, 2000 and 10000, for disease model 6 and scenario 4.



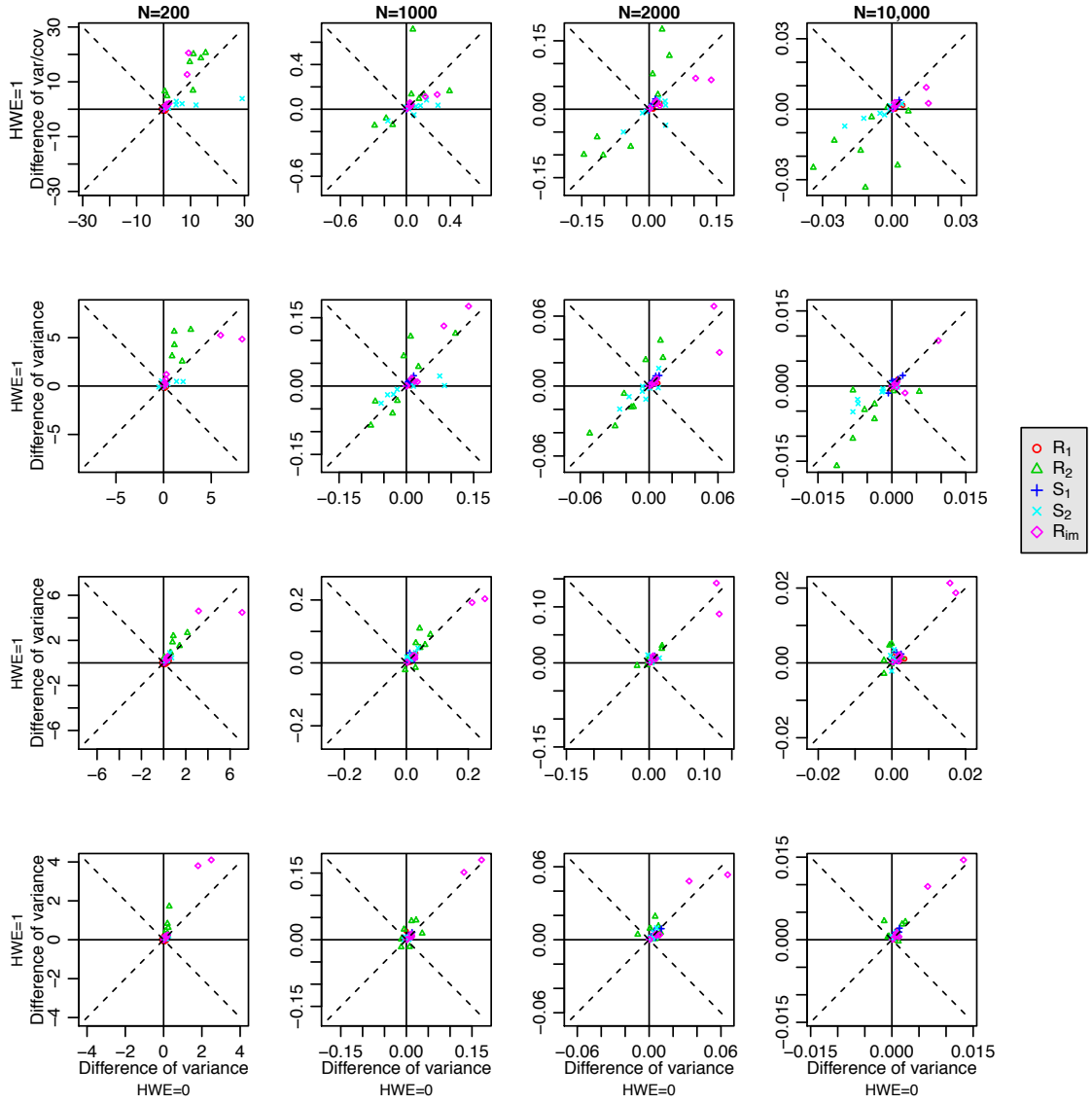
Supplementary Figure S8: QQ-norm plots for relative distance of a parameter and its LIME estimator with varying number of nuclear families: N = 200, 1000, 2000 and 10000, for disease model 7 and scenario 6.



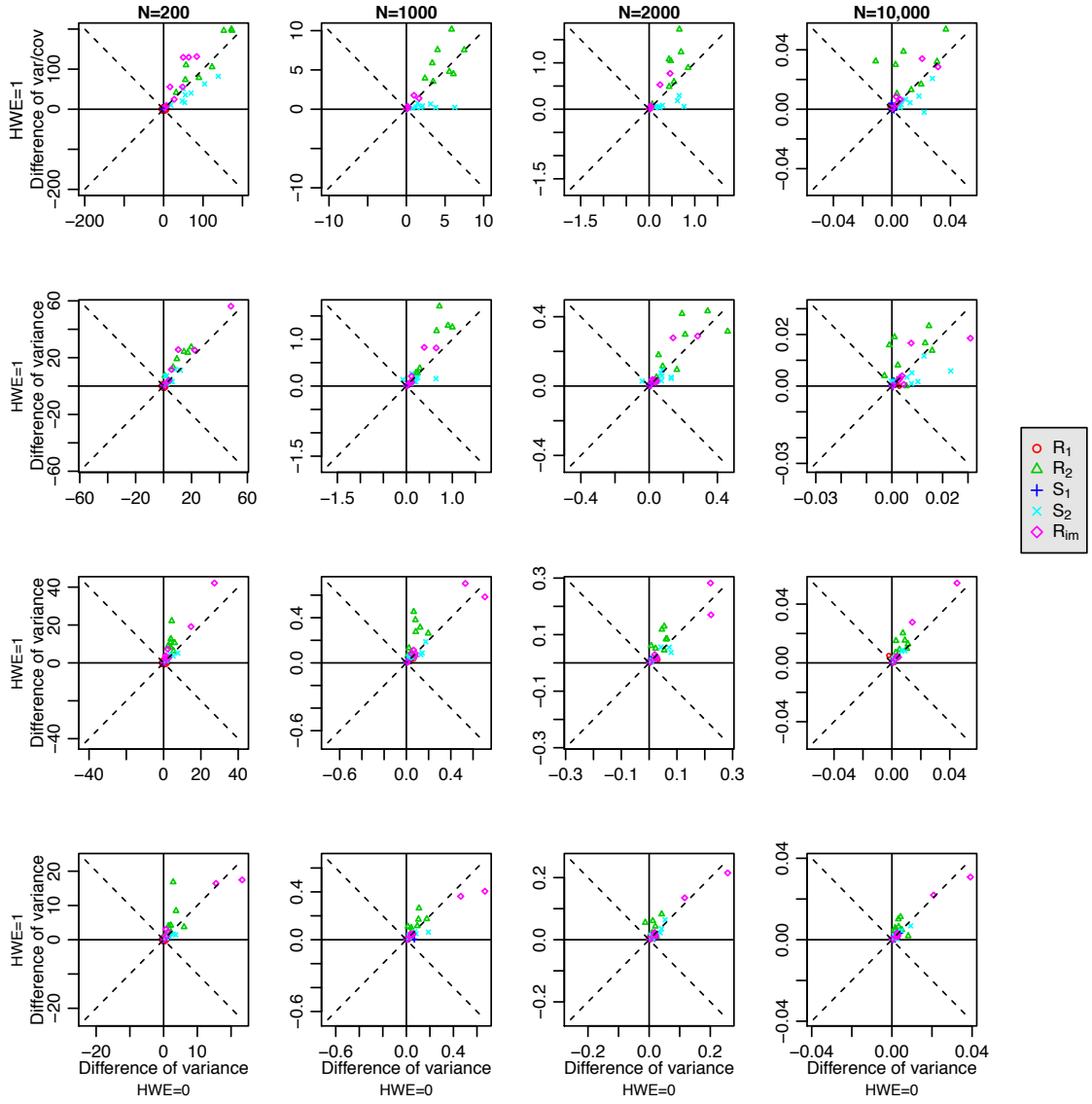
Supplementary Figure S9: The difference between empirical and asymptotic variances of estimators of $(R_1, R_2, S_1, S_2, R_{im})$ for HWE$=$ 1 vs. HWE$=$ 0 for T samples. Row 1: MAF$=$ 0.1, PREV$=$ 0.05, Row 2: MAF$=$ 0.1, PREV$=$ 0.15, Row 3: MAF$=$ 0.3, PREV$=$ 0.05 and Row 4: MAF$=$ 0.3, PREV$=$ 0.15.



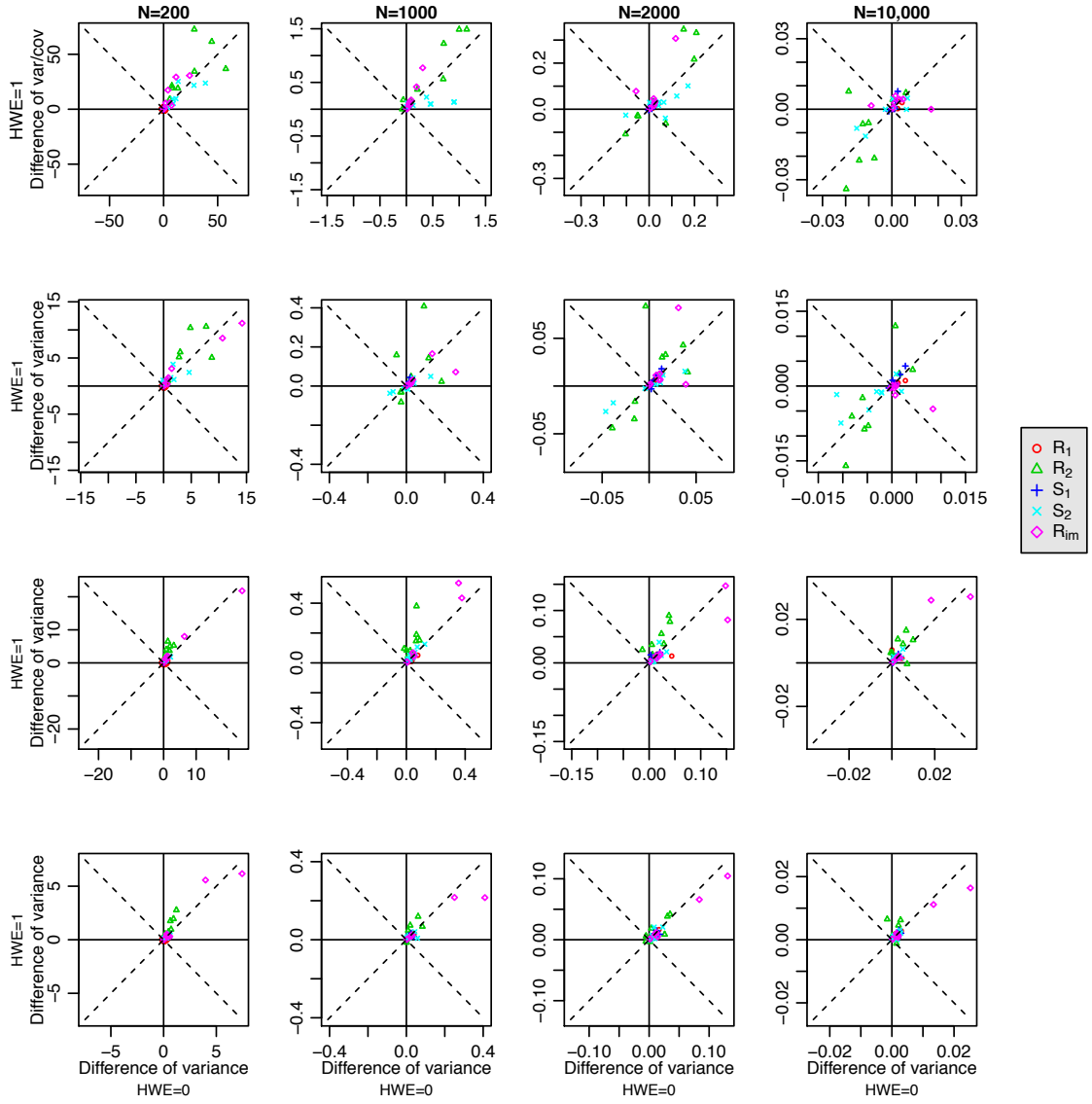
Supplementary Figure S10: The difference between empirical and asymptotic variances of estimators of $(R_1, R_2, S_1, S_2, R_{im})$ for HWE$=1$ vs. HWE$=0$ for $T+1$ samples. Row 1: MAF$=0.1$, PREV$=0.05$, Row 2: MAF$=0.1$, PREV$=0.15$, Row 3: MAF$=0.3$, PREV$=0.05$ and Row 4: MAF$=0.3$, PREV$=0.15$.



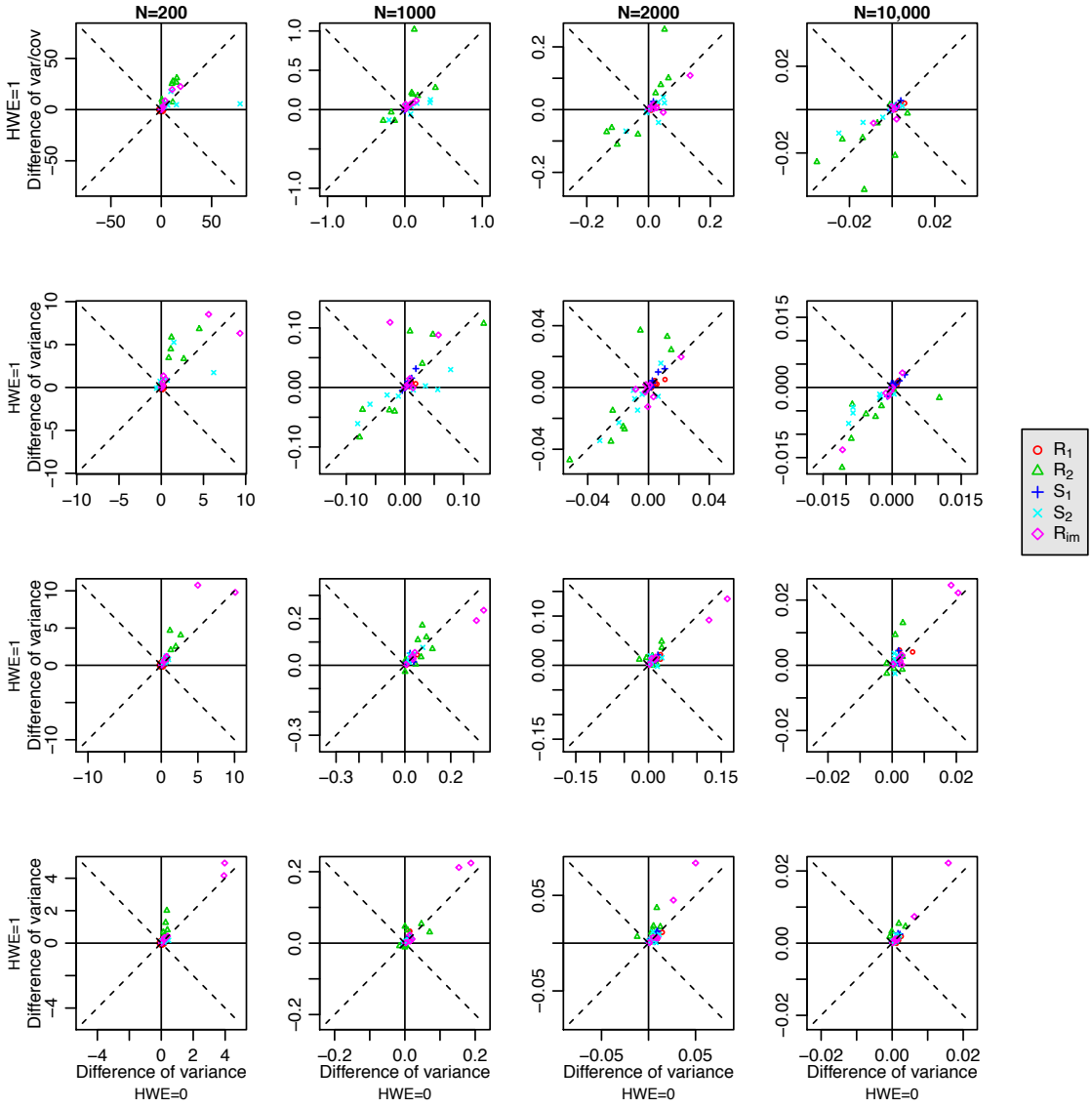
Supplementary Figure S11: The difference between empirical and asymptotic variances of estimators of $(R_1, R_2, S_1, S_2, R_{im})$ for HWE$=1$ vs. HWE$=0$ for $T+2$ samples. Row 1: MAF$=0.1$, PREV$=0.05$, Row 2: MAF$=0.1$, PREV$=0.15$, Row 3: MAF$=0.3$, PREV$=0.05$ and Row 4: MAF$=0.3$, PREV$=0.15$.



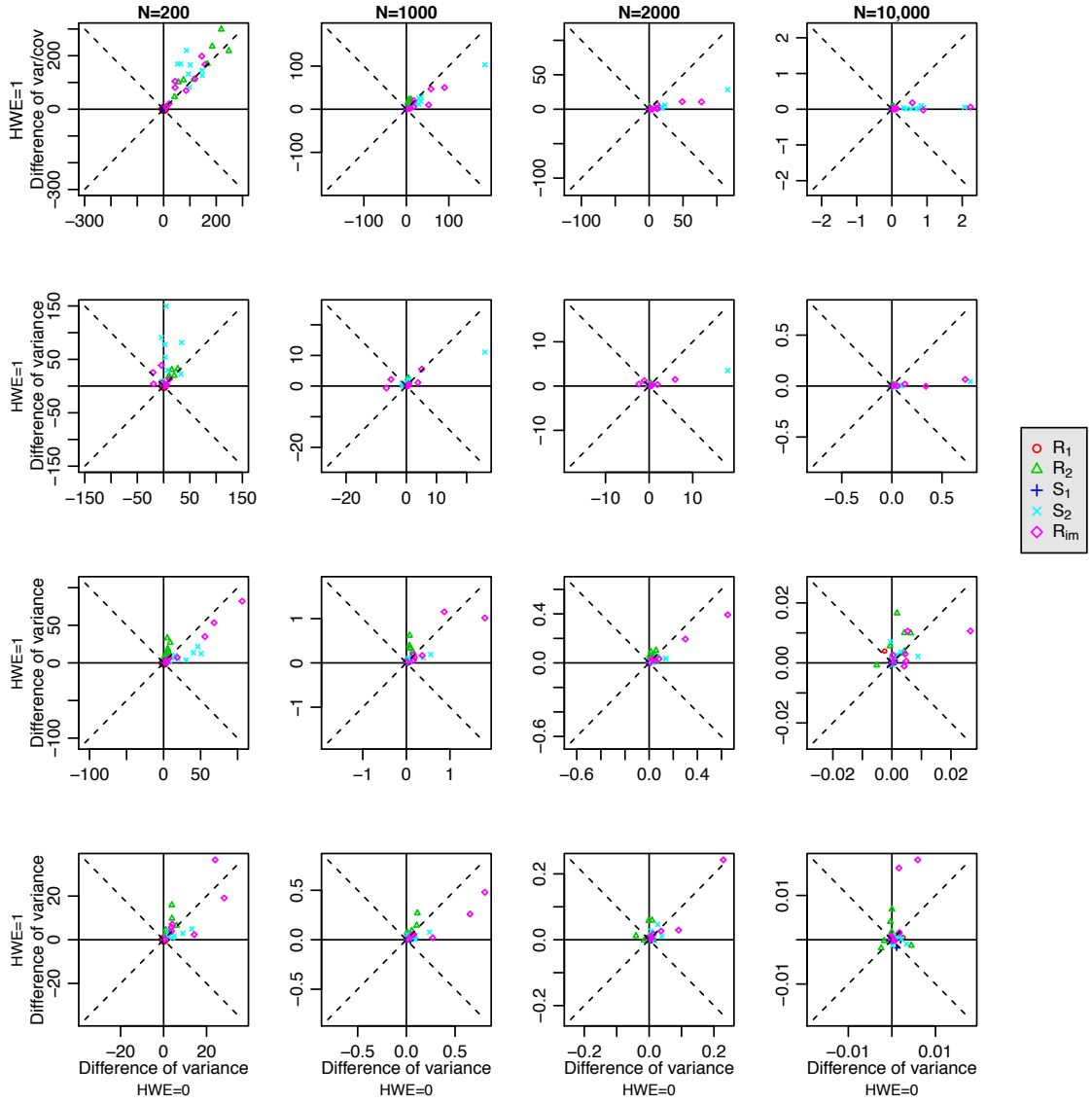
Supplementary Figure S12: The difference between empirical and asymptotic variances of estimators of $(R_1, R_2, S_1, S_2, R_{im})$ for HWE$=1$ vs. HWE$=0$ for M samples. Row 1: MAF$=0.1$, PREV$=0.05$, Row 2: MAF$=0.1$, PREV$=0.15$, Row 3: MAF$=0.3$, PREV$=0.05$ and Row 4: MAF$=0.3$, PREV$=0.15$.



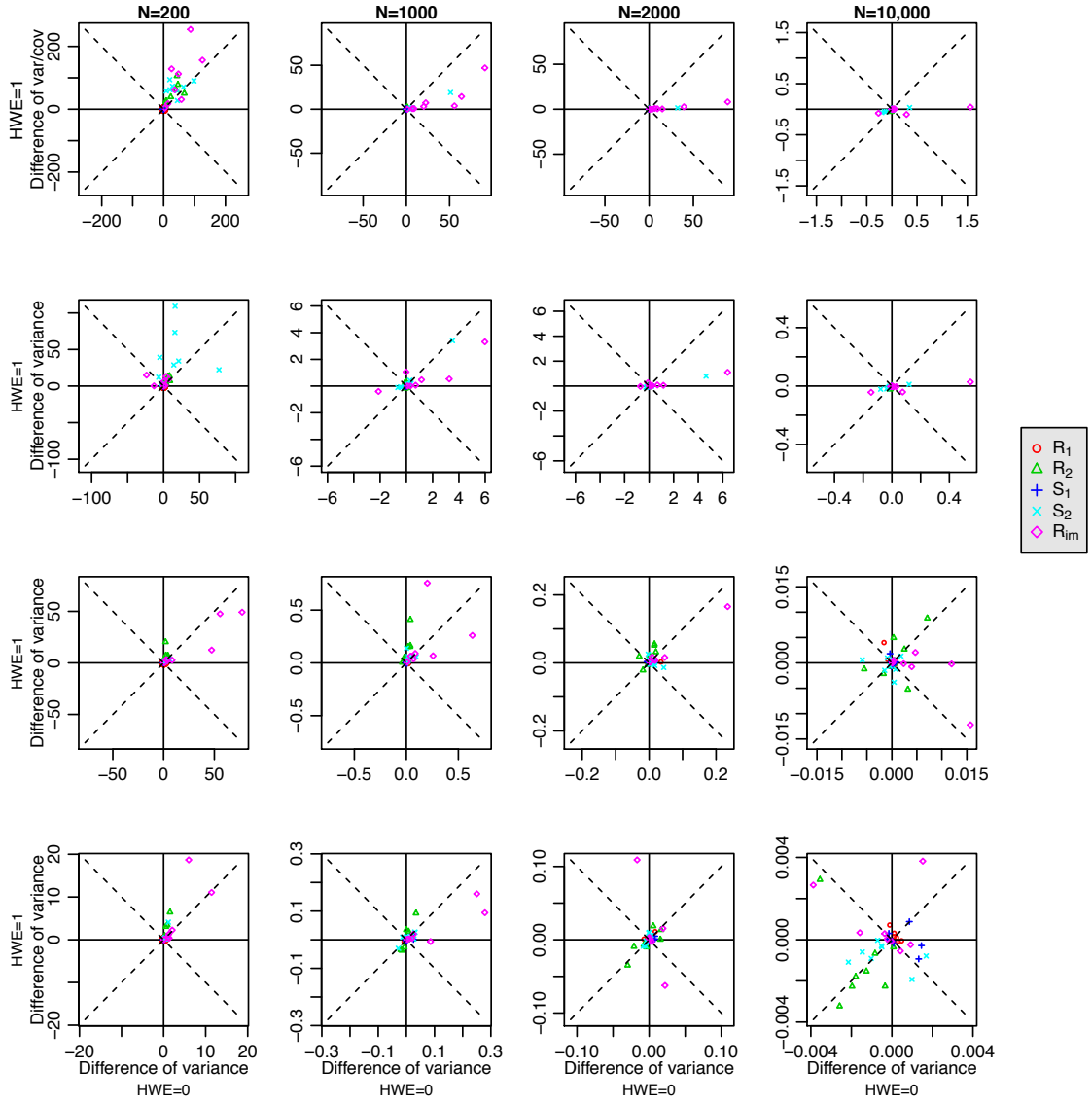
Supplementary Figure S13: The difference between empirical and asymptotic variances of estimators of $(R_1, R_2, S_1, S_2, R_{im})$ for HWE$=1$ vs. HWE$=0$ for $M+1$ samples. Row 1: MAF$=0.1$, PREV$=0.05$, Row 2: MAF$=0.1$, PREV$=0.15$, Row 3: MAF$=0.3$, PREV$=0.05$ and Row 4: MAF$=0.3$, PREV$=0.15$.



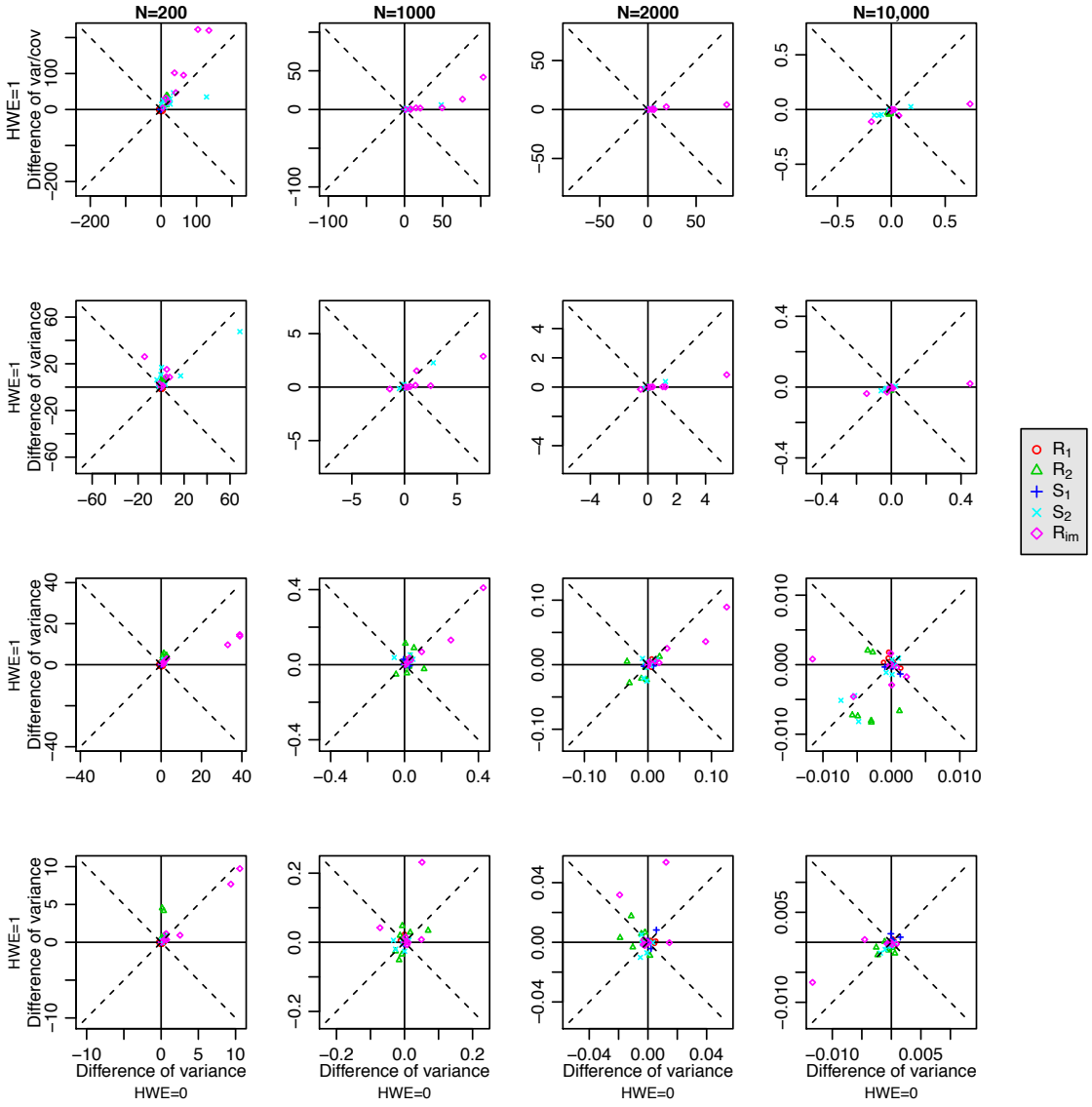
Supplementary Figure S14: The difference between empirical and asymptotic variances of estimators of $(R_1, R_2, S_1, S_2, R_{im})$ for HWE$=1$ vs. HWE$=0$ for $M+2$ samples. Row 1: MAF$=0.1$, PREV$=0.05$, Row 2: MAF$=0.1$, PREV$=0.15$, Row 3: MAF$=0.3$, PREV$=0.05$ and Row 4: MAF$=0.3$, PREV$=0.15$.



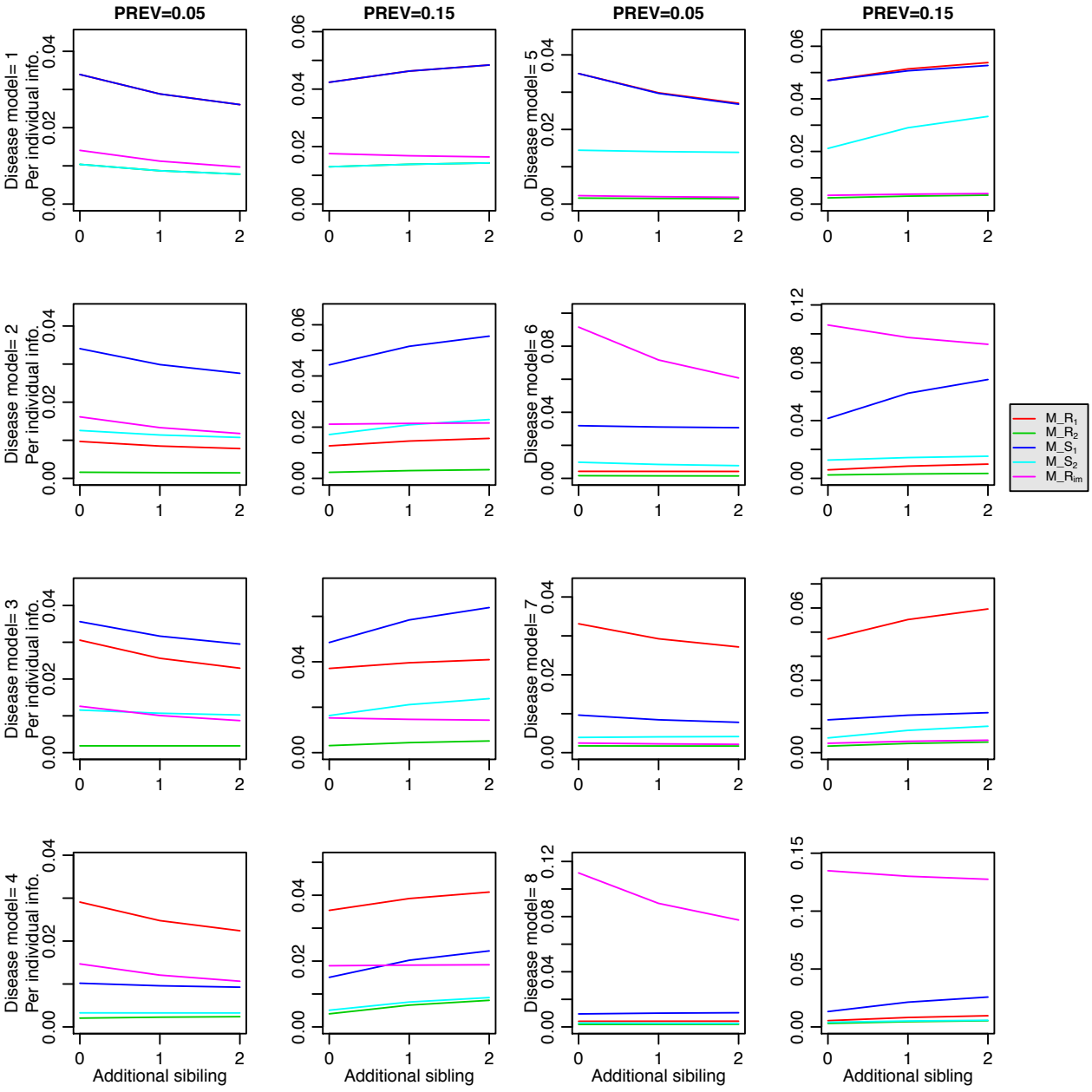
Supplementary Figure S15: The difference between empirical and asymptotic variances of estimators of $(R_1, R_2, S_1, S_2, R_{im})$ for HWE$=1$ vs. HWE$=0$ for P samples. Row 1: MAF$=0.1$, PREV$=0.05$, Row 2: MAF$=0.1$, PREV$=0.15$, Row 3: MAF$=0.3$, PREV$=0.05$ and Row 4: MAF$=0.3$, PREV$=0.15$.



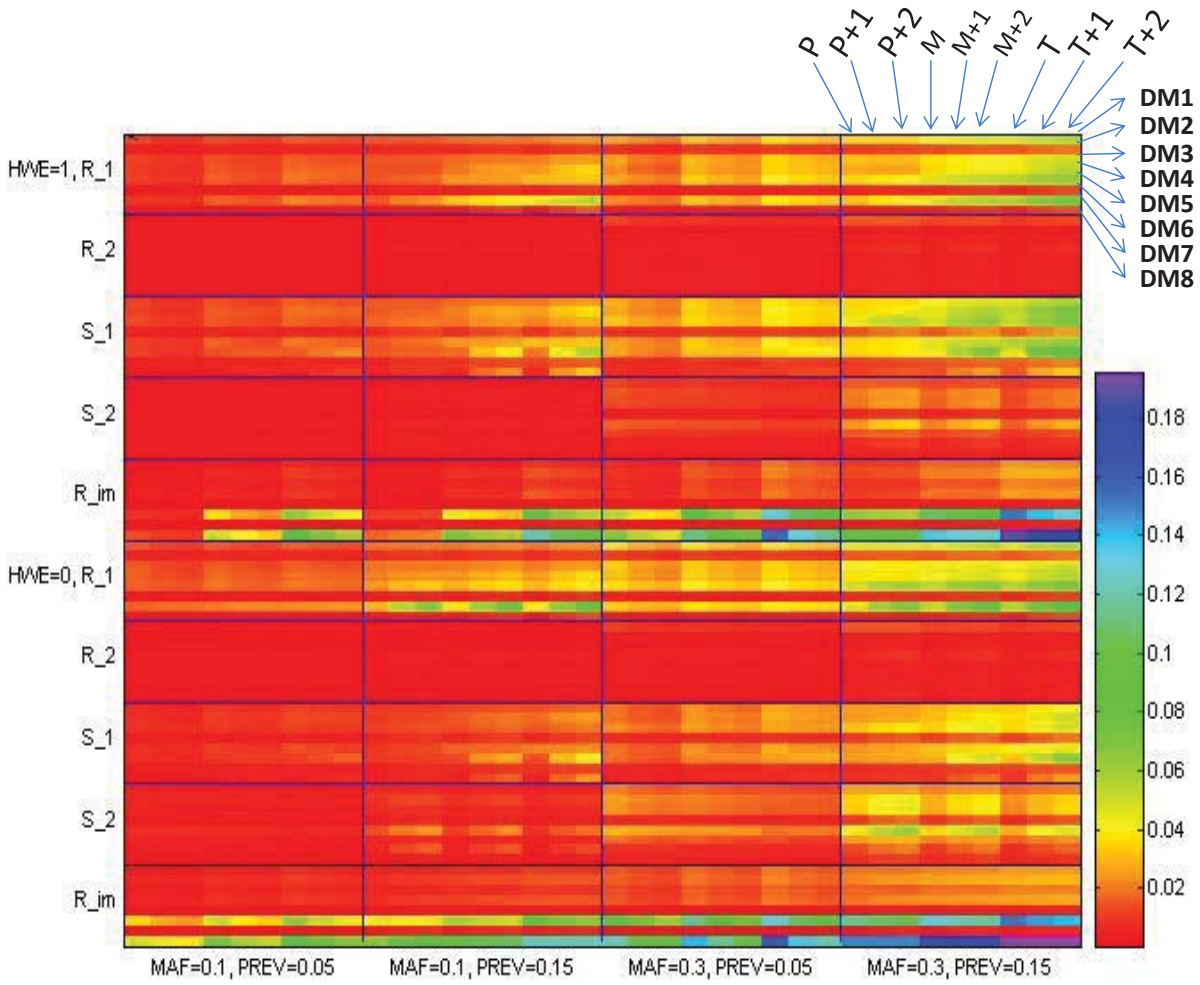
Supplementary Figure S16: The difference between empirical and asymptotic variances of estimators of $(R_1, R_2, S_1, S_2, R_{im})$ for HWE$=1$ vs. HWE$=0$ for $P+1$ samples. Row 1: MAF$=0.1$, PREV$=0.05$, Row 2: MAF$=0.1$, PREV$=0.15$, Row 3: MAF$=0.3$, PREV$=0.05$ and Row 4: MAF$=0.3$, PREV$=0.15$.



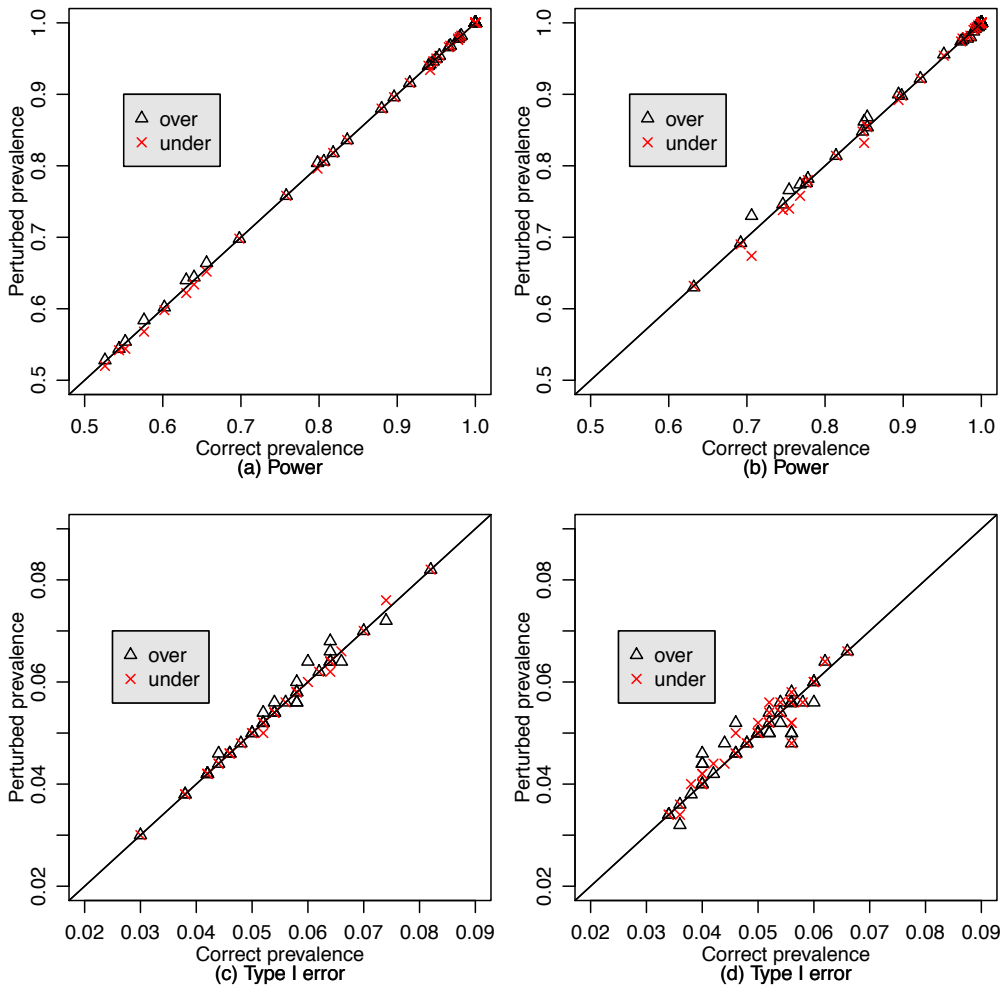
Supplementary Figure S17: The difference between empirical and asymptotic variances of estimators $(R_1, R_2, S_1, S_2, R_{im})$ for HWE$=1$ vs. HWE$=0$ for $P+2$ samples. Row 1: MAF$=0.1$, PREV$=0.05$, Row 2: MAF$=0.1$, PREV$=0.15$, Row 3: MAF$=0.3$, PREV$=0.05$ and Row 4: MAF$=0.3$, PREV$=0.15$.



Supplementary Figure S18: Information content per individual for 8 disease models and two PREVs with HWE $= 1$ and MAF $= 0.3$. Each curve provides the information for estimating one of the 5 parameters for the M data type without, with 1, or with 2 additional siblings.



Supplementary Figure S19: Per individual information content for parameter estimation from the 9 study designs (data types): $\{P,\ P+1,\ P+2,\ M,\ M+1,\ M+2,\ T,\ T+1,\ T+2\}$. Each of the 4 column blocks represent a MAF, PREV combination. Within each block, information from the 9 data types are presented in the order as indicated in the figure. In the top half, each of the 5 blocks provide information for the estimation of each of the 5 parameters under HWE. Furthermore, each of the 8 rows within a row block represents the 8 models. The bottom half provide the same information but with the HWE assumption being violated.



Supplementary Figure S20: Effects of mis-specification of prevalence on power (a and b) and type I error (c and d) of LIME. We considered perturbed prevalences that are 20% over, or 20% under, the true prevalence. The left plots are for true PREV $= 0.05$ and the right plots are for PREV $= 0.15$, over all eight disease models and eight scenarios described in Table 1 of the main text. The results show that mis-specification in the range investigated do not have a significant impact on power or type I error.