 DOI: 10.1111/j.1541-0420.2009.01270.x

A Likelihood-Based Trait-Model-Free Approach for Linkage Detection of Binary Trait

S. Basu,[1,*] M. Stephens,[2] J. S. Pankow,[3] and E. A. Thompson[4]

[1]Division of Biostatistics, University of Minnesota, 420 Delaware Street SE, Minneapolis, Minnesota 55455, U.S.A.
[2]Department of Statistics, University of Chicago, 5734 S. University Avenue, Chicago, Illinois 60637, U.S.A.
[3]Division of Epidemiology and Community Health, University of Minnesota, 1300 S. Second Street, Suite 300 Minneapolis, Minnesota 55454, U.S.A.
[4]Department of Statistics, University of Washington, B 313 Padelford Hall, Box 354322, Seattle, Washington 98195, U.S.A.
[*]*email:* saonli@umn.edu

Summary. Trait-model-free (or "allele-sharing") approach to linkage analysis is a popular tool in genetic mapping of complex traits, because of the absence of explicit assumptions about the underlying mode of inheritance of the trait. The likelihood framework introduced by Kong and Cox (1997, *American Journal of Human Genetics* **61,** 1179–1188) allows calculation of accurate p-values and LOD scores to test for linkage between a genomic region and a trait. Their method relies on the specification of a model for the trait-dependent segregation of marker alleles at a genomic region linked to the trait. Here we propose a new such model that is motivated by the desire to extract as much information as possible from extended pedigrees containing data from individuals related over several generations. However, our model is also applicable to smaller pedigrees, and has some attractive features compared with existing models (Kong and Cox, 1997), including the fact that it incorporates information on both affected and unaffected individuals. We illustrate the proposed model on simulated and real data, and compare its performance with the existing approach (Kong and Cox, 1997). The proposed approach is implemented in the program `lm_ibdtests` within the framework of `MORGAN 2.8` (`http://www.stat.washington.edu/thompson/Genepi/MORGAN/Morgan.shtml`).

Key words: Identity by descent; Likelihood ratio test; Linkage analysis, Trait model free.

1. Introduction

Trait-model-free approaches to linkage analysis attempt to test for linkage between a marker (or set of markers) and a trait locus by examining whether meioses at the marker, in a pedigree of related individuals, follow Mendelian segregation independent of the affectation status of individuals. They are based on the idea that, in a region of the genome linked to a trait locus, affected relatives are expected to share more alleles identical by descent (IBD) with other affected relatives and fewer alleles IBD with unaffected relatives than in regions unlinked to a trait locus.

There are two broad approaches to performing this kind of test. The first is based on directly specifying a test statistic that measures the extent to which affected relatives share alleles IBD (Whittemore and Halpern, 1994; Kruglyak et al., 1996). The second is based on specifying an explicit alternative model for the way that transmissions of alleles among individuals in the pedigree depend on the affectation status of the individuals, and performing a likelihood-ratio test against the null hypothesis that the transmission of alleles are independent of affectation status. A key advantage of this latter approach is that it is capable of dealing rigorously with the fact that IBD sharing is not observed directly, but must be inferred (with uncertainty) from observed marker data. Its use in affected-sibpair analyses was introduced by Risch (1990; see also Holmans, 1993). For larger pedigrees the first methods based on likelihood ratio tests were those introduced by Kong and Cox (1997).

In this article, we suggest a new explicit alternative model for transmission of alleles at a region of genome in the pedigree conditional on the affectation statuses of the relatives. This model is motivated by the idea that some founder chromosomes will carry a risk allele at the trait locus, and that these chromosomes will appear to be transmitted more frequently to affected relatives than unaffected relatives. In particular, the same founder chromosome may tend to appear together with the disease through multiple generations, and our model is aimed primarily at extracting this multigenerational information, which current models struggle to capture. However, it is applicable to pedigrees of any size, and has some attractive features, including that it does not require the specification of an IBD measure such as $S_{\text{pairs}}, S_{\text{all}}$ (Whittemore and Halpern, 1994; Kruglyak et al., 1996) and incorporates the data on unaffected individuals.

2. Methods

Let x denote the location in the genome of a putative trait locus. Following the notation of Kong and Cox (1997), we use ν to denote the inheritance vector (Lander and Green, 1987) at x. This can be thought of as a binary vector, with one

element for each meiosis in the pedigree, indicating whether (at x) the parent's maternal or paternal allele is transmitted to the offspring. Let Y denote observed marker data (typically at markers that span a region that includes x), and Φ the observed affectation statuses for pedigree members. Note that there may be many individuals whose affectation status and/or genotype data at markers are unknown.

Now consider testing the null hypothesis H_0 (that there is no linkage between the trait and location x) against the alternative hypothesis H_1 (that there is linkage between the trait and location x) using the marker data Y and the observed affectation statuses Φ. We assume that the distribution of Y depends on Φ, but only through ν. That is, if $\Pr_0$ and $\Pr_1$ denote probabilities under the null and alternative hypotheses, respectively, then $\Pr_1(Y \mid \Phi, \nu) = \Pr_1(Y \mid \nu) = \Pr_0(Y \mid \nu)$.

The likelihood ratio in favor of H_1 can then be written as

$$\begin{aligned}
\mathrm{LR} &= \frac{\Pr_1(Y \mid \Phi)}{\Pr_0(Y \mid \Phi)} \\
&= \mathrm{E}\left(\frac{\Pr_1(Y, \nu \mid \Phi)}{\Pr_0(Y, \nu \mid \Phi)}\middle| Y\right) \quad \text{(Thompson and Guo, 1991)} \\
&= \mathrm{E}\left(\frac{\Pr_1(\nu \mid \Phi)\Pr_0(Y \mid \nu)}{\Pr_0(\nu \mid \Phi)\Pr_0(Y \mid \nu)}\middle| Y\right) \\
&= \mathrm{E}\left(\frac{\Pr_1(\nu \mid \Phi)}{\Pr_0(\nu \mid \Phi)}\middle| Y\right), \qquad (1)
\end{aligned}$$

where $\mathrm{E}(.|Y)$ denotes expectation over the conditional distribution of ν given Y under H_0.

Under the null hypothesis, ν and Φ are independent, and assuming Mendelian inheritance $\Pr_0(\nu \mid \Phi) = (1/2)^N$, where N is the total number of meioses across all pedigrees. Under the alternative hypothesis, suppose that ν given Φ has a specified parametric form, with parameter vector δ: $\Pr_1(\nu \mid \Phi) = f(\nu; \Phi, \delta)$. Substituting these into equation (1) we obtain the following expression for the generalized likelihood ratio test statistic

$$\Lambda = 2\log(\mathrm{LR}) = 2\sup_{\delta}\log(\mathrm{E}(2^N f(\nu; \Phi, \delta) \mid Y)). \qquad (2)$$

A test of H_0 can be performed by first computing equation (2), and then computing a p-value by comparing the observed value with its (asymptotic) distribution under H_0. Computation of (2) can be performed exactly for pedigrees of moderate size (Kruglyak and Lander, 1998). For larger pedigrees it can be approximated, for example, using Markov chain Monte Carlo (MCMC) to sample from $\Pr_0(\nu \mid Y)$ as in Heath (1997), and using

$$\Lambda \approx 2\sup_{\delta}\log\left((1/M)\sum_i 2^N f(\omega_i; \Phi, \delta)\right), \qquad (3)$$

where $\omega_1, \ldots, \omega_M$ are MCMC samples from $\Pr_0(\nu \mid Y)$. Note that whether one performs the computation exactly, or by MCMC, it is necessary to assume a model for founder alleles. The standard assumptions in this setting are Hardy–Weinberg equilibrium and linkage equilibrium in the founders, and marker allele frequencies are known (Kruglyak et al., 1996; Thompson and Heath, 1999). Although these assumptions may not hold in practice, they are almost universally assumed in this setting, and we will make them here. We will return to the question of the asymptotic distribution of Λ under H_0 for our particular choice of f below.

The power of this test will clearly depend on the choice of f. Kong and Cox (1997) consider two possibilities such as a linear model (4) and an exponential model (5), each with a single scalar parameter δ:

$$f(\omega; \Phi, \delta) = (1/2)^N \prod_{i=1}^{n}(1 + \delta\gamma_i Z_i(\omega, \Phi)), \qquad (4)$$

and

$$f(\omega; \Phi, \delta) = (1/2)^N \exp\left(\delta\sum_i \gamma_i Z_i(\omega, \Phi)\right) \Big/ E_0\left[\exp\left(\sum_i \gamma_i Z_i(\omega, \Phi)\right)\right], \qquad (5)$$

where γ_i is a weighting factor for pedigree i whose value depends on the pedigree structure, and $Z_i = [S_i - E(S_i)]/\sqrt{(Var(S_i))}$, where S_i is a statistic (evaluated for pedigree i) chosen so that large values of S_i are evidence for linkage. To give just a simple example, S_i could be the total number of alleles shared IBD by every pair of affected relatives in pedigree i. The power of these models to detect linkage depends on choice of the IBD-sharing statistic Z_i and the weighting factors γ_i. Different choices perform better under different trait models (McPeek, 1999).

Here we suggest a new possible choice of f, which has two parameters: $\delta = (\lambda_a, \lambda_u)$. We begin by describing how to simulate from the distribution $f(\omega; \lambda_a, \lambda_u)$:

(1) Label the maternal and paternal allele in each founder with a 1 or 0 with probability 0.5, independently for each allele in each founder.

(2) Start at the top of the pedigree (i.e., with the founders), and use the assigned labels, together with the affectation status of individuals in the pedigree to simulate meioses from parent to offspring down to the bottom of the pedigree according to the following rules:

   (i) if the paternal and maternal alleles of the parent have the same label (i.e., both 0 or both 1), or if the affectation status of the offspring is unknown, transmit the paternal and maternal alleles with equal probability (0.5).

   (ii) if the paternal and maternal alleles of the parent have different labels (i.e., one labeled 0, the other labeled 1), and the offspring is *affected*, transmit the allele labeled 1 with probability λ_a, and the allele labeled 0 with probability $1 - \lambda_a$.

   (iii) if the paternal and maternal alleles of the parent have different labels, and the offspring is *unaffected*, transmit the allele labeled 1 with probability λ_u, and the allele labeled 0 with probability $1 - \lambda_u$.

Note that as an allele is transmitted down through the pedigree it takes its label with it. So that, for example, if a father passes an allele labeled 1 to his offspring then the paternal allele of the offspring is then labeled 1.

The intuition here is that alleles labeled 1 are nonbeneficial allele, and thus more likely to have been transmitted to affected offspring, while alleles labeled 0 are beneficial alleles, and more likely to have been transmitted to unaffected offspring. For this reason we impose the constraints $\lambda_a \geqslant 0.5$ and $\lambda_u \leqslant 0.5$. This particular choice of f thus models the preferential transmission of the nonbeneficial allele to an affected and the beneficial allele to an unaffected offspring from a parent.

The above describes how to simulate from f. However, to compute equation (2) we need to be able to compute f for any given value of ω. This requires a sum over all possible allocations, $\mathcal{A}$, of labels (0 and 1) to founder maternal and paternal alleles. This sum $\mathcal{A}$ for pedigree i contains 2^{2f_i} terms, where f_i is the total number of founders in pedigree i. For the pedigrees considered in our examples, which contain relatively few founders, this sum can be evaluated directly, and this is the approach we used. However, for pedigrees with many founders this naive approach would become impractical. For large pedigrees containing no loops the sum can be evaluated more efficiently using peeling methods (Lander and Green, 1987; Sobel and Lange, 1996; Thompson and Heath, 1999).

We refer to the distribution $f(\omega;\, \lambda_a,\, \lambda_u)$ described above as the "preferential transmission model" or PTM. Testing the null hypothesis of no linkage between the trait and location x will be equivalent to test, $\mathrm{H}_0' : \lambda_a = \lambda_u = 0.5$ under our model. The null hypothesis H_0' will test if the segregation of founder alleles at x is Mendelian. The asymptotic distribution of the likelihood-ratio statistic (equation 2) under H_0' is a mixture of chi-square $p_2\chi_2^2 + p_1\chi_1^2 + p_0\chi_0^2$ (see equation A3 in Appendix). When the marker data are completely informative about ν (*complete data case*), each of the mixing proportions p_2 and p_0 reduces to $\frac{1}{4}$ (see Lemma 2 in Appendix). We estimate the information matrix in equation (A3) by simulating a single dataset under no linkage keeping the pedigree structures, marker data availability on the individuals, genetic map of the markers, and marker allele frequencies the same as in the observed dataset. The mixing proportions for the mixture of chi-square distribution can then be computed from the estimated information matrix. An alternative approach would be to do an exact test with the likelihood-ratio statistic (equation 2) by permuting the affectation statuses of siblings within each sibship as in Basu, Di, and Thompson (2008).

One of the key motivations for our PTM approach is to gain power by including unaffected individuals in the analysis. To assess the extent of this gain, we also conducted affected-only analyses under the PTM (PTM aff-only) by setting $\lambda_u = 0.5$. For this analysis we need to estimate only one parameter, and so the asymptotic distribution of the LR statistic becomes a mixture of chi-square, $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ (Self and Liang, 1987).

The motivation behind this new PTM approach is that it provides a way to model the inheritance vector ν directly. This model assigns each founder allele either one of the two labels, such as "beneficial" or "nonbeneficial" type, thereby essentially classifying all distinct founder alleles into two categories. This can be more informative than scoring the IBD sharing among a group of relatives for each founder allele, especially when multiple founder alleles, associated with the disease are segregating in a pedigree. Hence for relatively common diseases, this approach may have higher power than any IBD scoring approach. Our approach also provides a way to combine the information on preferential transmission of alleles across all the pedigrees by having a common parameter λ_a and λ_u. Moreover, it provides a sensible way to incorporate the unaffected people into the model. Unaffected people within a pedigree structure can be served as controls to the cases (affected people) to make the IBD measure robust against phenocopies. Penrose (1939) first recommended the use of unaffected siblings as controls and later on this idea was implemented by several researchers (Curtis, 1997; Risch and Teng, 1998). Our proposed model takes into account the excess sharing of founder alleles among affected and also incorporates lack of sharing between an affected and unaffected individual.

The parameters of PTM can be interpreted through the parameters of a traditional parametric linkage analysis approach. At a disease locus, the alleles D and d can be labeled as the "nonbeneficial" and "beneficial" allele. In traditional parametric linkage analysis, we generally assume known disease allele frequencies and we assume known penetrance parameters at the locus. The penetrance parameters f_2, f_1, and f_0 represent $\Pr[\text{affected} \mid DD], \Pr[\text{affected} \mid Dd]$, and $\Pr[\text{affected} \mid dd]$, respectively. The disease allele frequency under the parametric model contributes to the number of disease alleles that segregate in a pedigree. In our model, we sum over all possible allocations of labels for the founder alleles after assigning equal weight to each allocation. One can introduce a parameter p as the population frequency of the "nonbeneficial" allele in order to assign different weights to each allocation $\mathcal{A}$ and estimate it from the data, which will be equivalent to estimating the disease allele frequency at the disease locus. Our approach models the trait dependent segregation of allele from a parent to an affected or an unaffected offspring using parameters λ_a and λ_u, respectively. The parents who are informative for the estimation of the parameters λ_a and λ_u are the ones with genotype Dd. For a nuclear family with two parents with genotype data Y_{P1} and Y_{P2} and an affected offspring with phenotype data Φ_o, one can easily establish the relationship between the penetrance parameters of the traditional linkage analysis and λ_a of our model. For example, for a parent with genotypes (Dd, dd),

$$\begin{aligned}\lambda_a &= \Pr[(Dd) \to D \mid \Phi_o = 1, (Y_{P1}, Y_{P2}) = (Dd, dd)] \\ &= \frac{\left[\Pr[(Dd) \to D, (dd) \to d \mid (Dd, dd)]\Pr[\Phi_o = 1 \mid Y_o = Dd]\right]}{\Pr[\Phi_o = 1 \mid (Dd, dd)]} \\ &= \frac{f_1}{f_0 + f_1}. \end{aligned} \tag{6}$$

Similarly, it can be shown that for the parents with genotypes (Dd, Dd) and an affected offspring, the relationship between λ_a and the penetrance parameters is $\lambda_a = \frac{f_1+f_2}{f_0+2f_1+f_2}$ and for parents with genotypes (Dd, DD), $\lambda_a = \frac{f_2}{f_1+f_2}$. The big difference between these two approaches is that our model conditions on the trait data and models $\Pr[\nu \mid \Phi]$, whereas traditional parametric linkage analysis models $\Pr[\Phi \mid \nu]$. Moreover, only the parents with one nonbeneficial and one beneficial allele are informative for our approach, which makes the parametric approach more powerful under the correctly specified trait model.

One issue with the current model is that it assumes Mendelian segregation of alleles from a parent to an offspring when the affectation status of the offspring is unknown. This can bias the findings toward the null hypothesis of no linkage if there is lot of missing phenotype data among the nonfounders in a pedigree. This problem can be addressed by imputing the relevant missing phenotypes from a distribution where each imputed phenotype receives a weight based on the number of affected and unaffected descendants of the individual. Then one can sum over all the imputed phenotypes to derive the final LR statistic. In this article, we have studied the performance of our model for two generation pedigrees with complete phenotypic information on the nonfounders and studied the performance of our model on a real dataset with three generation pedigrees with less than 20% missing phenotype data.

The likelihood ratio test under PTM is implemented in the `MORGAN` program `lm_ibdtests` released in `MORGAN version 2.8`. The current version of our likelihood-based approach allows to estimate the information matrix and provides an estimate of ρ specified in equation (A3) of the Appendix, which can then be used to compute the p-value of the observed likelihood-ratio statistic under the mixture of chi-square distribution. For every marker location, the program `lm_ibdtests` finally returns a p-value for linkage using the asymptotic mixture of chi-square distribution.

3. Results

3.1 *Simulation Study*

We considered 100 pedigrees each of size 6 for our simulation study. Each pedigree was a nuclear family with four offspring. We considered a diallelic marker with allele frequencies (0.5,0.5) for the simulation study. We assumed that the parents are unobserved for the trait and the marker data and the offspring are observed for both the marker and the trait data. Two offspring were affected and two were unaffected by the trait.

The program "markerdrop" in `MORGAN` (`http://www.stat.washington.edu/thompson/Genepi/pangaea.shtml`) was used for simulating marker data. Then the program "`lm_ibdtests`" in the `MORGAN` software was used to compute the likelihood-ratio statistic under PTM. We used the MCMC sampler of the `MORGAN` software (`http://www.stat.washington.edu/thompson/Genepi/MORGAN/Morgan.shtml`) for computation of the LR statistic under PTM. This sample pairs two block-Gibbs samplers, the locus (l) sampler of Heath (1997) and a meiosis (m) sampler to form the lm-sampler (Thompson and Heath, 1999). Note that, because we have only a single marker, we can obtain independent realizations from the conditional distribution of ν given Y. The MCMC lm-sampler of the `MORGAN` programs reduces, in this case, to independent sampling using the l-sampler of Heath (1997).

The program `Merlin` (Abecasis et al., 2002) was used to perform the LR test under the one parameter allele-sharing model (Kong and Cox, 1997). We used the IBD measure S_{pairs} (Whittemore and Halpern, 1994) and the linear model (equation 4) option in `Merlin` for our performance study. The other IBD measure S_{all} (Whittemore and Halpern, 1994) in `Merlin` is the same as S_{pairs} for affected sibpairs, hence we did not include S_{all} in our simulation study. We compared the performance of the one parameter allele-sharing model (Kong and Cox, 1997) with S_{pairs} with both PTM aff-only approach and PTM approach under various trait models. We are going to refer to the one parameter linear allele-sharing model (Kong and Cox, 1997) as the KC δ-model.

We first simulated marker data under the null hypothesis of no linkage between the trait and the marker. We calculated the statistic Λ (equation 2) for each dataset. Figure 1 shows the cumulative distributive function (cdf) of the Λ obtained empirically, on the basis of 5000 datasets simulated under the null hypothesis. We also estimated ρ (equation A3 in Appendix) from each dataset. The average ρ was 0.313 with a standard deviation of 0.018. We used the average ρ to calculate p_0 and p_2 in equation (A3). The cdf of the mixture of chi-square with the mixing proportions as stated in equation (A3) is also plotted in Figure 1. A Kolmogorov–Smirnov test between the empirical cdf of the Λ and the cdf of the mixture of chi-square (Figure 1) resulted in a p-value of 0.54. We also estimated the null distribution of the LR statistic under the KC δ-model and the PTM aff-only approach using the same set of simulated marker data.

We next compared the performance of PTM with the KC δ-model based on S_{pairs} (Kong and Cox, 1997). The trait locus was considered diallelic with alleles D and d and corresponding allele frequencies 0.2 and 0.8, respectively. We considered

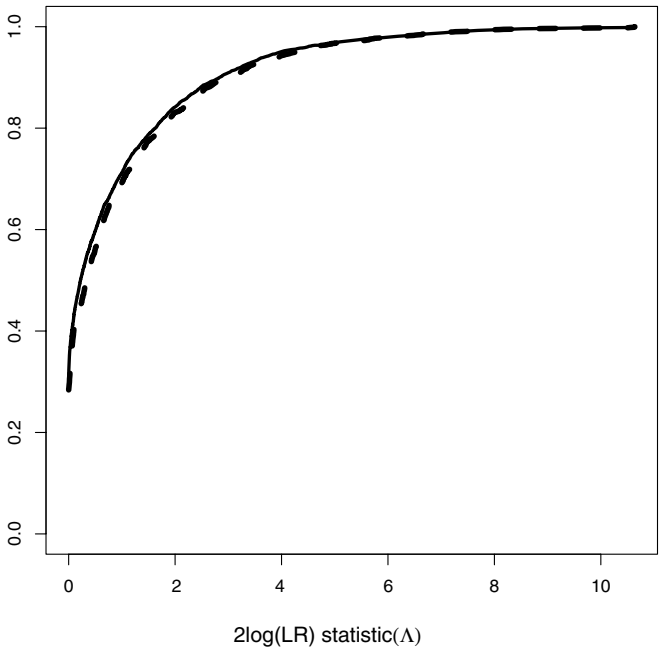

Figure 1. The figure shows the empirical distribution of the 2log (LR) test statistic (equation 2) for PTM under the null hypothesis. The solid line shows the cumulative distribution function of the mixture of chi-square $0.20\chi^2_2 + 0.50\chi^2_1 + 0.30\chi^2_0$. The proportions are computed using equation (A3) in Appendix. The dashed line shows the empirical cumulative distribution function of the 2 log (LR) statistic.

Table 1
Power of different approaches six different trait models with varying penetrances of the trait genotypes. The allele frequencies for the alleles D and d at the trait locus were 0.2 *and* 0.8, *respectively. Data were simulated at a marker completely linked to the trait. For each trait model, the power corresponding to a type I error of* 0.01 *are listed.*

Model	Penetrance	Power (PTM)	Power (PTM aff-only)	Power (KC δ model)
1	(0.05,0.95,0.95)	1.0	0.94	0.911
2	(0.05,0.50,0.95)	0.593	0.543	0.461
3	(0.05,0.05,0.95)	0.996	0.987	0.963
4	(0.15,0.85,0.95)	0.836	0.372	0.322
5	(0.15,0.65,0.99)	0.350	0.218	0.180
6	(0.15,0.15,0.99)	0.209	0.146	0.110

6 different trait models (Table 1) for this comparison study. The trait models were chosen to allow considerable amount of genetic heterogeneity and phenocopies in the data. Under different models in Table 1, one can calculate the probabilities of different genotypes at the disease locus given a person is affected or unaffected, using Bayes' theorem. For example, under model 4, the affected had 13% chance of being a *dd* genotype, and the unaffected people had 21% chance of carrying one copy of disease allele *D*. This can induce considerable genetic heterogeneity in the dataset, because several pedigrees can appear to be unlinked to the disease locus.

Marker datasets were simulated at a location completely linked to the trait locus. We simulated 1000 marker datasets under each trait model. For each simulated marker data, we ran `Merlin` and recorded the LR-statistic for linkage under the KC δ-model (Kong and Cox, 1997). On the same marker dataset, we ran `lm_ibdtests` to get the LR statistic under PTM and PTM aff-only model. The power to detect linkage for all these approaches was computed by calculating the proportion of LR statistics less than or equal to the 99th percentile of the distribution of the LR statistic under the null hypothesis. We have already shown in Figure 1 that the empirical quantiles for the LR statistic under PTM were very close to the quantiles of the mixture of chi-square distribution.

Table 1 shows the performance of the two approaches under six different trait models. The trait models were selected in such a way that a certain percentage of the affected people will carry the nondisease allele *d*, and a certain percent of unaffected people will carry the disease allele *D*. The dominant, additive, and recessive modes of inheritance of the trait were considered by choosing these six different trait models.

Under all six models, PTM had greater power than either PTM aff-only or the KC δ-model based on S_{pairs} (Table 1). The power gain was largest under model 4 (dominant with high phenocopy rate), where PTM had more than double the power of the other approaches. For all six models, the performance of PTM aff-only was slightly better than the KC δ-model (Table 1), suggesting that most of the power gain in the full PTM is due to the incorporation of information on unaffected individuals.

We also simulated 1000 marker datasets recombination fraction 0, 0.05, 0.1 for models 4, 5, and 6 of Table 1. The power is shown in Figure 2. For all three trait models, the PTM had substantially higher power to detect linkage at each recombination fraction, as compared to the PTM aff-only approach and the KC δ-model with the IBD measure S_{pairs}.

3.2 *Real Data Analysis*

The NHLBI Family Heart Study is a multicenter population-based study of genetic and nongenetic determinants of coronary heart disease, atherosclerosis, and cardiovascular risk factors. Families were enrolled at four U.S. centers (Higgins et al., 1996). We selected the continuous trait body mass index (BMI) measured at the baseline exam and dichotomized it in order to study the performance of different allele-sharing approaches. The main reason for choosing BMI was that a variance component analysis using `SOLAR` (Almasy and Blangero, 1998) on these pedigrees produced significant linkage signals on chromosomes 7 and 13 (Feitosa et al., 2002). A few other chromosomes also showed weak linkage signals. We dichotomized the trait using a cutoff value of 30, because a BMI above 30 is commonly used to define obesity (World Health Organization, 2000). All individuals with a BMI above 30 were assigned as affected and the rest of the individuals with known phenotypes were assigned as unaffected for our comparison study. Because there were missing phenotype data on some individuals, we decided to consider only the pedigrees with 20% or less missing phenotype data. We finally analyzed 30 pedigrees with 355 individuals. Among the pedigrees, 93% were three-generation pedigrees and the rest were two-generation pedigrees. The pedigrees had between 9 and 20 members. There were 295 people with known phenotypes (171 affected and 124 unaffected). There were 60 people with unknown phenotypes. We analyzed 396 microsatellite markers on 22 chromosomes with average heterozygosity of 75.1%.

We ran `Merlin` affected-only analysis with the IBD measures S_{pairs} and S_{all} using the KC δ-model. We then ran `lm_ibdtests` on the same set of pedigrees to compare the performance of our proposed model with the affected only analysis in `Merlin`. We also performed an affected-only analysis under our model. We used the MCMC sampler of the `MORGAN` software for computation of the LR statistic under both PTM and PTM aff-only approach. We estimated the correlation coefficient ρ for PTM by simulating data under the null hypothesis with similar marker data availability and marker allele frequencies. The average correlation coefficient was 0.05. In Figure 3, the p-values under our model were calculated using the null distribution of the LR statistic for complete data (Lemma 2), which is the most conservative test under our model. For all other approaches, we used the asymptotic chi-square approximation of the LR statistic and reported the asymptotic p-values. Figure 3 shows the performance of different allele-sharing approaches on this dataset.

We reported the findings for chromosomes 7 and 13, because they showed very significant evidence of linkage for the quantitative trait BMI (Feitosa et al., 2002). We also reported the chromosomes where at least one of the methods produced

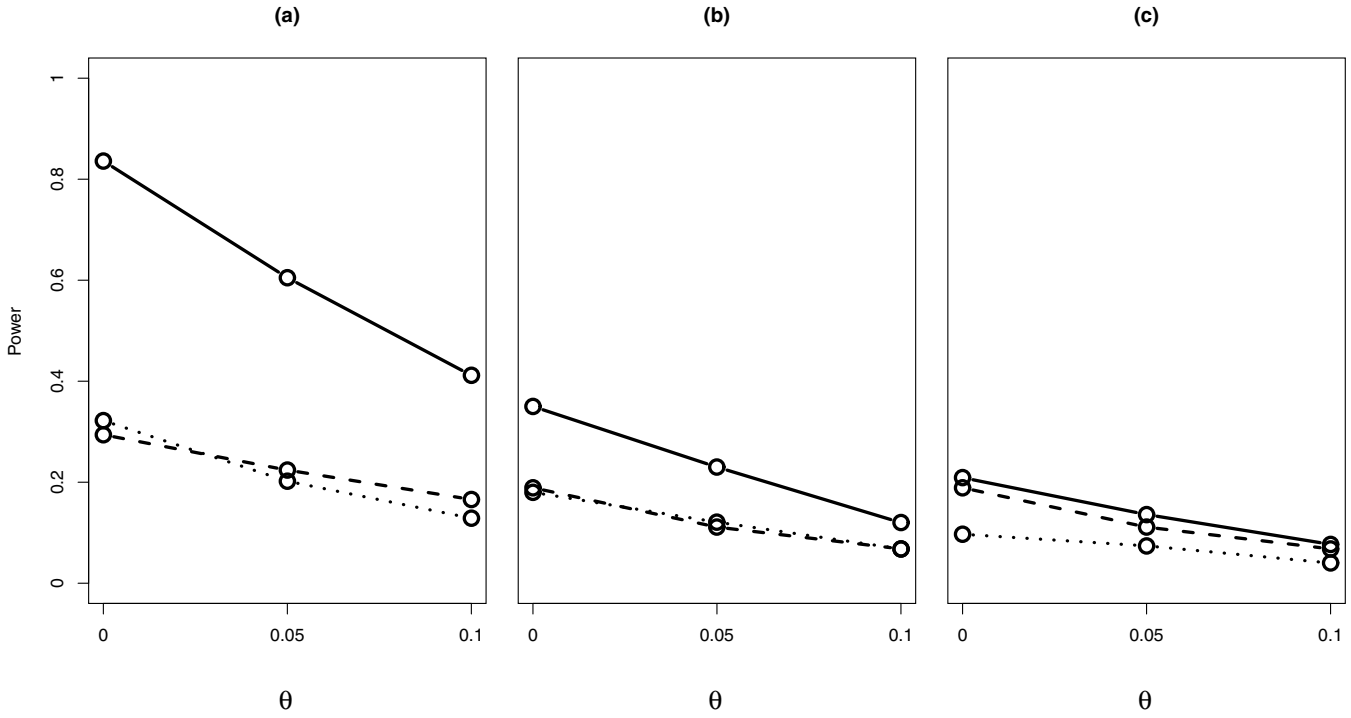

Figure 2. This figure shows the power for linkage detection of the likelihood-ratio test under PTM, PTM aff-only, and the one parameter linear allele-sharing model of Kong and Cox (1997) with S_{pairs} for (a) model 4, (b) model 5, and (c) model 6. The solid line in each graph shows the power to detect linkage at recombination fractions 0, 0.05, 0.1 under PTM. The dashed line and the dotted line in each graph show the power at different recombination fractions for PTM aff-only and for KC δ-model with S_{pairs}, respectively.

a linkage signal with a p-value lower than 0.001. The findings of the analysis incorporating both affected and unaffected people under our model resembled closely the findings from the variance component analysis using all pedigrees (Feitosa et al., 2002). The findings from the affected-only analysis were similar and they were somewhat different compared to the PTM analysis incorporating both affected and unaffected.

4. Discussion

The new likelihood-based approach models the trait dependent segregation of marker alleles due to linkage. It not only considers the excess IBD sharing among affected individuals due to linkage, it also takes into account the lack of sharing between affected and unaffected individuals. Moreover, the estimation of the parameters are carried out jointly using the data on multiple pedigrees. Hence, this model efficiently combines the data on multiple pedigrees to infer linkage. This two parameter model also avoids the use of any IBD measure. For all the trait models considered in this article, the proposed approach performed better than the affected-only analysis using the existing one parameter allelle-sharing model Kong and Cox (1997) and the affected only analysis under this PTM.

One can add another parameter p as the probability of a founder allele to be a nonbeneficial allele and estimate p from the available data. One issue of having p in the model is that the nonidentifiability of p under the null hypothesis complicates the asymptotic distribution of the likelihood-ratio statistic. Moreover, the information for p comes from the data on founders. If the data on founders are mostly missing, this additional parameter does not have much influence on the likelihood. In case the data have information on founders, maximizing p using the data will probably be more informative than fixing p at 0.5. We have investigated the power of such a model under few pedigree structures and marker data availability, but did not find any substantial gain in power by adding this third parameter.

The current model assumes Mendelian segregation of founder alleles for offspring with unknown phenotype. This can be a potential problem if there is a lot of missing phenotype data for the nonfounders. This problem can be addressed by imputing the relevant missing phenotype based on the number of affected and unaffected descendants of an individual and calculating a weighted LR statistic by assigning a weight to each imputed disease status. We intend to further investigate this issue.

The model has similarities with the gamete competition models of Sinsheimer, Blangero, and Lange (2000), where alleles in a parent are considered to compete with one another for transmission to offspring. Their model is aimed toward detecting association of alleles with a trait, and so the probability that an allele is transmitted depends on its allelic state. In contrast, our model is aimed toward detecting linkage, and so the transmission probability depends on which founder allele it is descended from. Just as Sinsheimer et al. (2000)

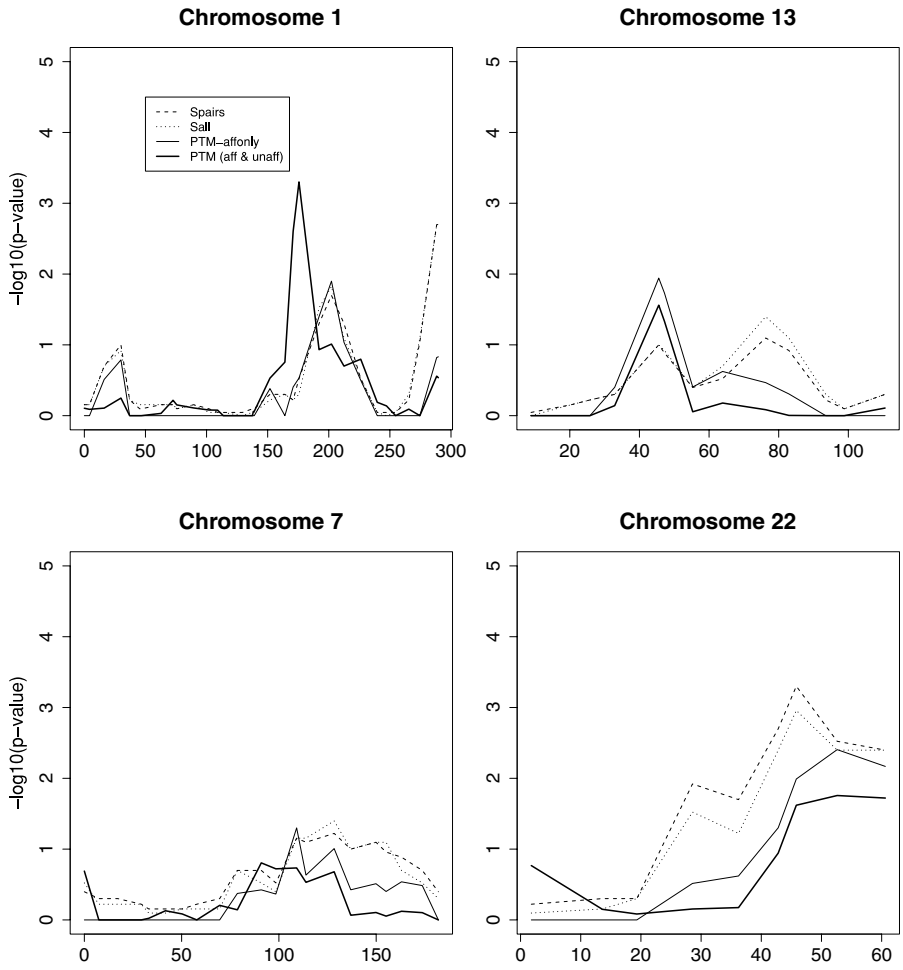

Figure 3. This figure shows the performance of the likelihood-ratio test under PTM (dark solid line), PTM aff-only (solid line), and the one parameter linear allele-sharing model of Kong and Cox (1997) for (a) S_{pairs} (dashed line), (b) S_{all} (dotted line) for linkage detection of obesity trait in NHLBI Family Heart Study. We have reported the linkage signals on chromosomes 1, 7, 13, and 22.

have a different parameter for each allele, one could imagine having a different parameter for each founder allele in our approach. However, this would create a large number of parameters, which would have a negative impact on precision of estimates, and potentially on power. Conversely one could imagine adapting our approach (with just two risk classes, and a parameter p indicating the proportion of alleles in each risk class) to the association context, and it might be interesting to compare this with Sinsheimer et al.'s (2000) approach.

Acknowledgements

This research was supported by PMMB grant and NIH grant GM-46255. The authors are thankful to Myrna Jewett for her contribution in developing the program `lm_ibdtests` and to Minnesota Supercomputer Institute for using their resources for processing and management of the Family Heart Study dataset. The authors would also like to thank the two anonymous referees for their helpful comments.

References

Abecasis, G., Cherny, S., Cookson, W., and Cardon, L. (2002). Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics* **30,** 97–101.

Almasy, L. and Blangero, J. (1998). Multipoint quantitative trait linkage analysis in general pedigrees. *American Journal of Human Genetics* **62,** 1198–1211.

Basu, S., Di, Y., and Thompson, E. A. (2008). Exact trait-model-free tests for linkage detection in pedigrees. *Annals of Human Genetics* **72,** 676–682.

Curtis, D. (1997). Use of siblings as controls in case-control association studies. *American Journal of Human Genetics* **61,** 319–333.

Feitosa, M. F., Borecki, I. B., Rich, S., Arnett, D. K., Sholinsky, P., Myers, R. H., Leppert, M., and Province, M. A. (2002). Quantitative-trait loci influencing body-mass index reside on chromosomes 7 and 13: The national heart, lung, and blood institute, Family Heart Study. *American Journal of Human Genetics* **70,** 72–82.

Heath, S. C. (1997). Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *American Journal of Human Genetics* **61,** 748–760.

Higgins, M., Province, M., Heiss, G., Eckfeldt, J., Ellison, R. C., Folsom, A. R., Rao, D. C., Sprafka, J. M., and Williams, R. (1996). NHLBI Family Heart Study: Objectives and design. *American Journal of Epidemiology* **143,** 1219–1228.

Holmans, P. (1993). Asymptotic properties of affected-sib-pair linkage analysis. *American Journal of Human Genetics* **52**, 362–374.

Kong, A. and Cox, N. J. (1997). Allele-sharing models: LOD scores and accurate linkage tests. *American Journal of Human Genetics* **61,** 1179–1188.

Kruglyak, L. and Lander, E. S. (1998). Faster multipoint linkage analysis using Fourier transformation. *Journal of Computational Biology* **5,** 1–7.

Kruglyak, L., Daly, M. J., Reeve-Daly, M. P., and Lander, E. S. (1996). Parametric and nonparametric linkage analysis: A unified multipoint approach. *American Journal of Human Genetics* **58,** 1347–1363.

Lander, E. S. and Green, P. (1987). Construction of multilocus genetic maps in humans. *Proceedings of National Academy of Sciences USA* **84,** 2363–2367.

McPeek, S. (1999). Optimal allele-sharing statistics for genetic mapping using affected relatives. *Genetic Epidemiology* **16,** 225–249.

Penrose, L. (1939). Some practical considerations in testing for genetic linkage in sib data. *Ohio Journal of Science* **39,** 291–296.

Risch, N. (1990). Linkage strategies for genetically complex traits. I. multilocus models. *American Journal of Human Genetics* **46,** 222–228.

Risch, N. and Teng, J. (1998). The relative power of family-based and case-control designs for association studies of complex human diseases. I. DNA pooling. *Genome Research* **8,** 1273–1288.

Self, S. G. and Liang, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of American Statistical Association* **82,** 605–610.

Sinsheimer, J. S., Blangero, J., and Lange, K. (2000). Gamete-competition models. *American Journal of Human Genetics* **66,** 1168–1172.

Sobel, E. and Lange, K. (1996). Descent graphs in pedigree analysis: Applications to haplotyping, location scores, and marker-sharing statistics. *American Journal of Human Genetics* **58,** 1323–1337.

Thompson, E. A. and Guo, S. W. (1991). Evaluation of likelihood ratios for complex genetic models. *Mathematical Medicine and Biology* **8,** 149–169.

Thompson, E. A. and Heath, S. C. (1999). Estimation of conditional gene identity among relatives. In *Statistics in Molecular Biology and Genetics: Selected Proceedings of a 1997 Joint AMS-IMS-SIAM Summer Conference on Statistics in Molecular Biology, IMS Lecture Note-Monograph Series*, F. Sellier-Moiseiwitch (ed), 95–113. Hayward, California: Institute of Mathematical Statistics.

Whittemore, A. S. and Halpern, J. (1994). A class of tests for linkage using affected pedigree members. *Biometrics* **50,** 118–127.

World Health Organization. (2000). *Obesity: Preventing and managing the global epidemic.* Technical Report 894 (ISBN 92-4-120894-5), Geneva: World Health Organization.

Received June 2008. *Revised January* 2009. *Accepted February* 2009.

Appendix

Likelihood for Complete Data

Consider a specific risk allocation $\mathcal{A}$ of the founder alleles in all n pedigrees. We define by $\mathrm{N}_a(\mathcal{A})$ as the total number of meioses that involve transmission of allele from a parent to an affected offspring, where the parent has alleles of type (0,1). Let $\mathrm{N}_u(\mathcal{A})$ be the total number of meioses that involve transmission of allele from a parent to an unaffected offspring, where the parent has alleles of type (0,1). We define by $\mathrm{X}(\mathcal{A})$ as the number of meioses where an affected offspring received an allele of type 1 from a parent with alleles (0,1) for a specific allocation $\mathcal{A}$. Then $\mathrm{X}(\mathcal{A})$ is a binomial with parameters $\mathrm{N}_a(\mathcal{A})$ and λ_a. Similarly, we define $\mathrm{Y}(\mathcal{A})$ as the number of meioses where an unaffected offspring received an allele of type 1 from a parent with alleles (0,1) for a specific allocation $\mathcal{A}$. Then $\mathrm{Y}(\mathcal{A})$ is distributed as binomial with parameters $\mathrm{N}_u(\mathcal{A})$ and λ_u. The other possibility could be that in a meiosis the affectation status of the involved offspring will be unknown or the parent will have alleles of type (0, 0) or type (1,1). The transmission probability of an allele from the parent to the offspring will then be $\frac{1}{2}$ under our model.

With the above defined variables, the likelihood for complete data takes the form:

$$\begin{aligned}\mathrm{P}[\nu=\omega\,|\,\Phi] &= \sum_{\mathcal{A}}\left(\prod_{i=1}^{n}\mathrm{P}[\omega_i\,|\,\mathcal{A}]\right)P[\mathcal{A}]\\ &= \sum_{\mathcal{A}}\Bigg[(\lambda_a)^{X(\mathcal{A})}(1-\lambda_a)^{(N_a(\mathcal{A})-X(\mathcal{A}))}(\lambda_u)^{Y(\mathcal{A})}\\ &\qquad\times(1-\lambda_u)^{(N_u(\mathcal{A})-Y(\mathcal{A}))}\\ &\qquad\times\left(\frac{1}{2}\right)^{(N-N_a(\mathcal{A})-N_u(\mathcal{A}))}\Bigg]P[\mathcal{A}]. \qquad \text{(A1)}\end{aligned}$$

A.1 *Asymptotic Distribution of LR Statistic*

If $\mathrm{L}(\delta)$ denotes $\mathrm{Pr}_1[Y\mid\Phi]$, where Y is the observed marker data that do not provide complete information on segregation of alleles (ν) at location x, then for $j=1,\,2,\,3$

$$\begin{aligned}\dot{l}_j(\delta_0) &= \left.\frac{\partial\log(\mathrm{L}(\delta))}{\partial\delta_j}\right|_{\delta_0}\\ &= \left.\frac{1}{\mathrm{L}(\delta_0)}\frac{\partial L(\delta)}{\partial\delta_j}\right|_{\delta_0}\\ &= \left.\frac{1}{\mathrm{Pr}_0(Y)}\frac{\partial}{\partial\delta_j}\left(\sum_{\nu}\mathrm{Pr}_1[\nu\,|\,\Phi]\,\mathrm{Pr}[Y\,|\,\nu]\right)\right|_{\delta_0}\\ &= \left.\frac{\partial}{\partial\delta_j}\left[\mathrm{E}\left(\frac{\mathrm{Pr}_1(\nu\,|\,\Phi)}{\mathrm{Pr}_0[\nu]}\middle|Y\right)\right]\right|_{\delta_0}\\ &= \left.\frac{\partial}{\partial\delta_j}\big(\mathrm{E}\big(2^N\,\mathrm{Pr}_1(\nu\,|\,\Phi)\big|Y\big)\big)\right|_{\delta_0}. \qquad \text{(A2)}\end{aligned}$$

The root-N consistency of the maximum likelihood estimator $\widehat{\delta}$ of $\delta=(\lambda_a\,,\lambda_u)$ for the constrained parameter space space of $\Omega=\{\lambda_a\in[0.5,\,1],\,\lambda_u\in[0,\,0.5]\}$ follows from theorem 1 of Self and Liang (1987). Following the same notation of Self and Liang (1987), we denote [0.5,1]$\times$ [0,0.5] by Ω, and $\Omega_0=\delta_0=(0.5,\,0.5)$. If $Z=\sqrt{n}((\widehat{\lambda}_a-0.5),(0.5-\widehat{\lambda}_u))$, then under the null hypothesis $Z=(Z_1,\,Z_2)\sim N_2(0,\,I^{-1}\,(\delta_0),I(\delta_0))$ is the information matrix. Lemma 1 shows that for incomplete data Y, the covariance of $Z=(Z_1,\,Z_2)$ is nonzero, but for complete data case, the estimators $(\widehat{\lambda}_a-0.5)$ and $(0.5-\widehat{\lambda}_u)$

are independently distributed, because $\mathrm{E}(-\ddot{l}_{12}(\delta_0)) = 0$ (Lemma 2). The true parameter belongs to the border of the Ω.

The asymptotic distribution of the likelihood-ratio statistic Λ then follows directly from the case 7 of Section 3 in Self and Liang (1987). Define $\rho = I_{12}/\sqrt{I_{11}}\sqrt{I_{22}}$, where $I(\delta_0) = (I_{11}, I_{12}, I_{21}, I_{22})$ is the information matrix of $\widehat{\lambda}_a$ and $\widehat{\lambda}_u$ computed at δ_0. Then under the null hypothesis,

$$\Lambda = 2\log(\mathrm{LR}) \sim p_0\chi_0^2 + p_1\chi_1^2 + p_2\chi_2^2,$$

where,

$$\begin{aligned} p_2 &= \frac{1}{2\pi}[cos^{-1}(\rho)], \\ p_0 &= \frac{1}{2\pi}[\pi - cos^{-1}(\rho)], \\ p_1 &= 1 - p_2 - p_0 = \frac{1}{2}, \quad \text{and} \end{aligned} \tag{A3}$$

Lemma 1: *Under the null hypothesis* H_0,

$$\mathrm{E}(-\ddot{l}_{12}(\delta_0)) < \infty, \tag{A4}$$

Proof.

$$\begin{aligned} \mathrm{E}(-\ddot{l}_{12}(\delta_0)) &= \mathrm{E}\left[-\frac{\partial^2 \log(\mathrm{L}(\delta)}{\partial\lambda_a\partial\lambda_u}\bigg|_{\delta_0}\right] \\ &= \left(\frac{1}{\mathrm{L}(\delta_0)}\right)^2 \mathrm{E}\left[\frac{\partial \mathrm{L}(\delta)}{\partial\lambda_a}\bigg|_{\delta_0}\frac{\partial \mathrm{L}(\delta)}{\partial\lambda_u}\bigg|_{\delta_0}\right] \\ &\quad - \mathrm{E}\left(\frac{1}{\mathrm{L}(\delta_0)}\frac{\partial^2 \mathrm{L}(\delta))}{\partial\lambda_a\partial\lambda_u}\bigg|_{\delta_0}\right) \\ &= \mathrm{E}\left[\frac{\partial \log(\mathrm{L}(\delta))}{\partial\lambda_a}\bigg|_{\delta_0}\frac{\partial \log(\mathrm{L}(\delta))}{\partial\lambda_u}\bigg|_{\delta_0}\right] \\ &\quad - \mathrm{E}\left(\frac{1}{\mathrm{L}(\delta_0)}\frac{\partial^2 \mathrm{L}(\delta))}{\partial\lambda_a\partial\lambda_u}\bigg|_{\delta_0}\right) \\ &= \mathrm{E}\left[\frac{\partial \log(\mathrm{L}(\delta))}{\partial\lambda_a}\bigg|_{\delta_0}\frac{\partial \log(\mathrm{L}(\delta))}{\partial\lambda_u}\bigg|_{\delta_0}\right] \\ &\quad - \mathrm{E}\left(\frac{1}{\mathrm{L}(\delta_0)}\frac{\partial^2}{\partial\lambda_a\partial\lambda_u}\left(\mathrm{E}\left(2^N \Pr_1(\nu \mid \Phi) \mid Y\right)\right)\right)\bigg|_{\delta_0} \\ &= \mathrm{E}\left[\frac{\partial \log(\mathrm{L}(\delta))}{\partial\lambda_a}\bigg|_{\delta_0}\frac{\partial \log(\mathrm{L}(\delta))}{\partial\lambda_u}\bigg|_{\delta_0}\right] \\ &\quad - \frac{1}{\mathrm{L}(\delta_0)}\frac{\partial^2}{\partial\lambda_a\partial\lambda_u}\left(\mathrm{E}\left(\mathrm{E}\left(2^N \Pr_1(\nu \mid \Phi) \mid Y\right)\right)\right)\big|_{\delta_0} \\ &= \mathrm{E}\left[\frac{\partial \log(\mathrm{L}(\delta))}{\partial\lambda_a}\bigg|_{\delta_0}\frac{\partial \log(\mathrm{L}(\delta))}{\partial\lambda_u}\bigg|_{\delta_0}\right] \\ &\quad - \frac{1}{\mathrm{L}(\delta_0)}\frac{\partial^2}{\partial\lambda_a\partial\lambda_u}\left(2^N\left(\frac{1}{2}\right)^N \sum_{\nu}(\Pr_1(\nu \mid \Phi))\right)\bigg|_{\delta_0} \\ &= 2^{2N}\left[\mathrm{E}\left(\mathrm{E}\left(\frac{\partial \Pr_1[\nu \mid \Phi]}{\partial\lambda_a}\bigg|_{\delta_0}\bigg| Y\right)\right.\right. \\ &\quad \left.\left.\times \mathrm{E}\left(\frac{\partial \Pr_1[\nu \mid \Phi]}{\partial\lambda_u}\bigg|_{\delta_0}\bigg| Y\right)\right)\right] - 0 \\ &= K < \infty. \end{aligned} \tag{A5}$$

Lemma 2: *For complete data case,* $\mathrm{E}(-\ddot{l}_{12}(\delta_0)) = 0$.

Proof. For the complete data case, the equation reduces to

$$\mathrm{E}(-\ddot{l}_{12}(\delta_0)) = 2^{2N}\left[\mathrm{E}\left(\left(\frac{\partial \Pr_1[\nu \mid \Phi]}{\partial\lambda_a}\bigg|_{\delta_0}\right)\left(\frac{\partial \Pr_1[\nu \mid \Phi]}{\partial\lambda_u}\bigg|_{\delta_0}\right)\right)\right].$$

Using the form of the complete data log likelihood in equation (A1), the above expression will have the form,

$$\begin{aligned} &\mathrm{E}(-\ddot{l}_{12}(\delta_0)) \\ &= 2^{2N} \times \left(\frac{1}{2}\right)^{2N} \mathrm{E}\left[\left(\sum_{\mathcal{A}} 2(2X(\mathcal{A}) - N_a(\mathcal{A}))\Pr[\mathcal{A}]\right)\right. \\ &\qquad \left.\times \left(\sum_{\mathcal{A}'} 2(2Y(\mathcal{A}') - N_u(\mathcal{A}'))\Pr[\mathcal{A}']\right)\right] \\ &= 4\sum_{\mathcal{A}}\sum_{\mathcal{A}'} \Pr[\mathcal{A}]\Pr[\mathcal{A}']\mathrm{E}[(2X(\mathcal{A}) - N_a(\mathcal{A})) \\ &\qquad \times (2Y(\mathcal{A}') - N_u(\mathcal{A}'))] \\ &= 4\sum_{\mathcal{A}}\sum_{\mathcal{A}'} \Pr[\mathcal{A}]\Pr[\mathcal{A}']\mathrm{E}[\mathrm{E}((2X(\mathcal{A}) - N_a(\mathcal{A})) \\ &\qquad \times (2Y(\mathcal{A}') - N_u(\mathcal{A}') \mid N_a(\mathcal{A}), \\ &\qquad N_u(\mathcal{A}'))], \end{aligned}$$

where the inside expectation is w.r.t $\Pr[\nu \mid N_a(\mathcal{A}), N_u(\mathcal{A}')]$

$$\begin{aligned} &= 4\sum_{\mathcal{A}}\sum_{\mathcal{A}'} \Pr[\mathcal{A}]\Pr[\mathcal{A}']\mathrm{E}[(\mathrm{E}(2X(\mathcal{A}) - N_a(\mathcal{A})) \mid N_a(\mathcal{A})) \\ &\qquad \times (\mathrm{E}(2Y(\mathcal{A}') - N_u(\mathcal{A}') \mid N_u(\mathcal{A}'))], \end{aligned}$$

because given $N_a(\mathcal{A})$ and $N_u(\mathcal{A}')$, $X(\mathcal{A})$ and $Y(\mathcal{A})$ are independent

$$\begin{aligned} &= 0, \qquad \text{because } \mathrm{E}(X(\mathcal{A}) \mid N_a(\mathcal{A})) = \frac{N_a(\mathcal{A})}{2} \text{ and} \\ &\qquad \mathrm{E}(Y(\mathcal{A}') \mid N_u(\mathcal{A}')) = \frac{N_u(\mathcal{A}')}{2}. \end{aligned} \tag{A6}$$