Allowing for Missing Parents in Genetic Studies of Case-Parent Triads

C. R. Weinberg

National Institute of Environmental Health Sciences, Research Triangle Park, NC

Summary

In earlier work, my colleagues and I described a log-linear model for genetic data from triads composed of affected probands and their parents. This model allows detection of and discrimination between effects of an inherited haplotype versus effects of the maternal haplotype, which presumably would be mediated by prenatal factors. Like the transmission disequilibrium test (TDT), the likelihood-ratio test (LRT) based on this model is not sensitive to associations that are due to genetic admixture. When used as a method for testing for linkage disequilibrium, the LRT can be regarded as an alternative to the TDT. When one or both parents are missing, the resulting incomplete triad must be discarded to ensure validity of the TDT, thereby sacrificing information. By contrast, when the problem is set in a likelihood framework, the expectation-maximization algorithm allows the incomplete triads to contribute their information to the LRT without invalidation of the analysis. Simulations demonstrate that much of the lost statistical power can be recaptured by means of this missing-data technique. In fact, power is reasonably good even when no triad is complete—for example, when a study is designed to include only mothers of cases. Information from siblings also can be incorporated to further improve the statistical power when genetic data from parents or probands are missing.

Introduction

Parents offer a rich source of genetic information for inference related to the joint presence of linkage and association for a marker allele or a candidate allele (Falk and Rubinstein 1987; Self et al. 1991; Spielman et al. 1993; Spielman and Ewens 1996). Moreover, the use of parents as controls (rather than unrelated persons) inherently prevents the influence of etiologically meaningless associations that can exist in a population that is stratified owing to incomplete mixing of genetically distinct subpopulations.

For effects mediated through the inherited genotype, the transmission disequilibrium test (TDT) can detect the apparent nonrandomness in transmission to affected offspring that arises when a particular allele is associated with the disease and also is linked to (or is itself) a susceptibility gene (Spielman et al. 1993). For effects mediated through the maternal phenotype—for example, through prenatal mechanisms—the TDT applied to case-parent triads is of no use, because transmission to the child is unrelated to the child's disease status (Wilcox et al. 1998).

My colleagues and I have proposed a likelihood-ratio test (LRT; based on a 2-df χ^2 statistic) developed by maximizing the likelihood under a log-linear model that describes the multinomial distribution for case-parent triads, with stratification by parental mating type (Weinberg et al. 1998; Wilcox et al. 1998). The proposed method offers better power than the TDT (on the basis of simulations), under either a dominant or a recessive genetic model, and can be used to test for effects of the offspring haplotype, the maternal haplotype, or both simultaneously. The fact that under either a dominant or a recessive model the 2-df LRT tends to outperform the 1-df TDT may seem counterintuitive. The difference is best understood as arising because the LRT is based on a general model formulation and uses information about the joint transmission from the parents, rather than accounting for the parental transmissions separately. These points have been discussed elsewhere (Weinberg et al. 1998).

A practical problem with the use of parents as controls is that one or both parents may be unavailable for genetic study: parents may be dead or unavailable, they may refuse to participate, or the father may need to be excluded post hoc because he was identified incorrectly as the parent of the proband. The result is incomplete triads, perhaps with genetic data only for the affected individual and one parent (a dyad) or neither parent (a monad).

In some families, a dyad could, in theory, contribute

Received November 12, 1998; accepted for publication February 2, 1999; electronically published March 8, 1999.

Address for correspondence and reprints: Dr. C. R. Weinberg, MD A3-03, National Institute of Environmental Health Sciences, P.O. Box 12233, Research Triangle Park, NC 27709. E-mail: weinberg@niehs.nih.gov

0002-9297/99/6404-0031$02.00

information to the TDT, because transmission sometimes is implied by the genotype of the proband and that of the parent who provided data. For example, if the offspring is homozygous for the variant allele and the parent studied is heterozygous, then, although the missing parent cannot be used, we can infer that the parent studied must have transmitted the allele to the child. Other dyads would have to be discarded, because they would provide only ambiguous information. For example, when the child and parent are both heterozygous, we cannot know whether the copy carried by the child came from the parent studied or from the missing parent. Curtis and Sham (1995) have pointed out, however, that a TDT analysis that selectively excludes ambiguous dyads while including the unambiguous dyads is invalid, and, therefore, only complete triads should be used. Unfortunately, the restriction to complete triads discards information and can cause a substantial loss of statistical power.

The purpose of the present article is to describe a likelihood-based method for inclusion of genetic information from incomplete triads and to characterize the behavior of the proposed method, through simulations. Once the problem has been cast in a likelihood framework, powerful statistical techniques for the handling of missing data can be applied, and the partial information from dyads and monads can be exploited fully, without sacrificing the validity of the test.

Background

Assume that there is a single affected proband from each family studied, and assume for simplicity that the gene under consideration is biallelic or can be meaningfully split into two categories of alleles for analysis. Let *M, F,* and *C* denote the number of copies of the variant allele carried by the mother, the father, and the child, respectively. Case-parent triads can fall into any of 15 possible categories, as shown in table 1. Here, these categories are grouped by parental mating type, as in the article by Schaid and Sommer (1993), and symmetric matings are assumed to be equally likely (e.g., [$M = 2$, $F = 1$] and [$M = 1$, $F = 2$]). The third column of table 1 describes hypothetical frequencies for the 15 cells; R_1 (R_2) is the adjusted relative risk for a child with one copy (two copies) of the variant allele, compared with a child with no copies. Mendelian inheritance is assumed. Under a null hypothesis of no linkage, R_1 and R_2 are 1.0, and the multinomial distribution is specified only by the mating-type–stratum parameters μ_j, together with the assumption of Mendelian inheritance. If the genetic mechanism is through maternal effects rather than through effects of the inherited gene, then a similar table can be constructed, except that the relative-risk multipliers, designated S_1 and S_2, correspond to M rather than to C (see table 2 in Wilcox et al. 1998).

Table 1

Frequencies for Case-Parent Triads

M, F, C[a]	Mating Type	Theoretical Frequency[b]
2, 2, 2	1	$R_2\mu_1$
2, 1, 2	2	$R_2\mu_2$
2, 1, 1	2	$R_1\mu_2$
1, 2, 2	2	$R_2\mu_2$
1, 2, 1	2	$R_1\mu_2$
2, 0, 1	3	$R_1\mu_3$
0, 2, 1	3	$R_1\mu_3$
1, 1, 2	4	$R_2\mu_4$
1, 1, 1	4	$2R_1\mu_4$
1, 1, 0	4	μ_4
1, 0, 1	5	$R_1\mu_5$
1, 0, 0	5	μ_5
0, 1, 1	5	$R_1\mu_5$
0, 1, 0	5	μ_5
0, 0, 0	6	μ_6

[a] *M, F,* and *C* denote the number of copies of the variant allele carried by the mother, the father, and the child, respectively.

[b] R_1 and R_2 are the relative risks associated with inheriting one or two copies, respectively, of the variant allele; μ_j is the stratum parameter for the *j*th mating-type category.

The use of logarithms yields a linear model for the logged expected counts in the 15 cells, with the most general form of the model (Wilcox et al. 1998) specifying the log of the expected count as

$$\omega_j + \beta_1 I_{(C=1)} + \beta_2 I_{(C=2)} + \alpha_1 I_{(M=1)} + \alpha_2 I_{(M=2)}$$

$$\omega_j + \ln(2) I_{(M=F=C=1)} \ . \qquad (1)$$

Here, $I_{(C=1)}$ denotes an indicator (0/1) variable that is set to 1 for cells where $C = 1$, etc. R_1 corresponds to $\exp(\beta_1)$, R_2 to $\exp(\beta_2)$, S_1 to $\exp(\alpha_1)$, and S_2 to $\exp(\alpha_2)$. The six mating-type–stratum parameters are included via ω_j.

The general model can be fit by use of any standard software for Poisson regression and can be reduced to impose an assumption that effects are mediated strictly via either the inherited C or the maternal M (my colleagues and I also have described an extension that allows for effects of parental imprinting [Weinberg et al. 1998], but this extension will not be considered here). Specific genetic alternatives can be imposed. For example, a dominant model is implied by constraining $\beta_1 = \beta_2$, which is easily accomplished by omission of $I_{(C=1)}$ and $I_{(C=2)}$ in favor of the composite $I_{(C>0)}$ variable. Alternatively, a recessive model can be specified by omission of the indicator variables for heterozygosity. This modeling approach can be regarded as a generalization of the approach suggested by Schaid and Sommer (1993), who also described maximum-likelihood methods conditional on the parental genotype.

There are important parallels between the LRT and the TDT. Under a strict null hypothesis that the allele under study is neither linked to nor associated with the disease, the parameters β_1 and β_2 in table 1 are 0, and the LRT statistic has a central 2-df χ^2 distribution, whereas the TDT statistic has a central 1-df χ^2 distribution. If there is association in the population but no linkage, the null distributions of both test statistics are unaffected. If there is linkage but no association, and, in contrast with the methods described by Ewens and Spielman (1995), random mating is not assumed for the parental generation, neither distribution will be χ^2, but both tests would have little power. Thus, equation (1) (and its corresponding likelihood) provides a method for statistically testing for linkage disequilibrium (via the β parameters), in that rejection of the null hypothesis suggests the presence of both association and linkage.

In earlier work, my colleagues and I performed simulations to study the operating characteristics of the 2-df χ^2 LRT of the null hypothesis that $\beta_1 = \beta_2 = 0$ (Weinberg et al. 1998). The testing was performed under the broad class of genetic alternatives, without constraining the alternative to be, for example, a dominant model. The simulated population was composed of two admixed subpopulations that had very different gene prevalences and different baseline risks of disease (risks among individuals who carry no copies of the allele). On the basis of those simulations, when both tests could be applied (i.e., the genetic mechanism was not maternally mediated), the LRT outperformed the TDT under either a recessive or a dominant genetic model but was slightly underpowered, compared with the TDT, under the gene-dose model in which $R_2 = R_1^2$.

Handling Missing Parents by Use of the Expectation-Maximization (EM) Algorithm

Now suppose that, for a given triad, the father is missing and $M = 2 = C$. This father must be either $F = 2$ or $F = 1$, but the value for F cannot be known. Assume that "missingness" is unrelated to (M, F, C) in that the probability that a father is missing is not related to the allele under study. If the parameters in table 1 are known, the probability that the father is in fact $F = 2$ is $\mu_1/(\mu_1 + \mu_2)$. Thus, roughly that proportion of dyads of the form $(M, F, C) = (2, \text{unknown}, 2)$ would have come from the first row of table 1. This is the idea underlying the statistical method to be applied.

The approach proposed is a standard and easily applied statistical method for handling missing data, called the "EM algorithm" (Dempster et al. 1977). In the present context, the method fractionally assigns the incomplete triads to their theoretically possible cells, on the basis of current estimates of the parameters (the E step), then repeats the maximization (the M step) of the likelihood on the basis of the newly revised, pseudocomplete data. The method then revises the provisional assignments of the incomplete triads (repeating the E step). This two-step alternating computation is repeated until convergence has been achieved. The likelihood then considered must be that based on the observed data (not the pseudocomplete data) but determined by use of the estimated parameters. The likelihood contribution for dyads is based on the sum of probabilities across the cells with which the observed data are compatible; in the example given, this likelihood contribution would be based on the sum across the first two cells of table 1. Details are provided in Appendix A. Statistical theory guarantees that the observed-data likelihood (including the partial information from dyads and monads) increases to a maximum via the algorithm; thus, an LRT can be performed validly (Dempster et al. 1977). My experience with simulated data has been that convergence under the EM is achieved reliably in <40 iterations, by means of this simple multinomial application of the method. Thus, simulations have become feasible.

Simulation Methods

As in earlier work (Weinberg et al. 1998), the simulations were set up on the basis of an admixture of two subpopulations. For a 20% subpopulation, the gene prevalence was .3, and the background risk was .05; for the remaining 80% of the population, the gene prevalence was .1, and the background risk was .01. For simplicity, each of the two subpopulations was assumed to be in Hardy-Weinberg equilibrium, even though the resulting mixed population was not and mating was not random in the mixed population. By this construction, there is a strong, positive (and etiologically meaningless) association between the allele and the disease in the population, even when $R_1 = R_2 = 1$. All simulations and analyses were performed by use of the GLIM package (Baker and Nelder 1978).

For each genetic scenario considered, 1,000 studies were simulated, each of which began with 100 case-parent triads. If a proportion p of 1,000 simulated studies reject the null hypothesis, then the empirical standard error for the rejection rate (size under the null hypothesis or power under alternatives to the null hypothesis) can be approximated by $\sqrt{p(1-p)/1{,}000}$. Approximate 95% confidence intervals for the true rejection rate then can be constructed by adding and subtracting twice this number to p. For simplicity, the initial simulations assumed that only the father could be missing, but this scenario was extended later to allow mothers to be missing as well. The missingness of fathers was assumed to be unrelated to the allele under study and was assigned

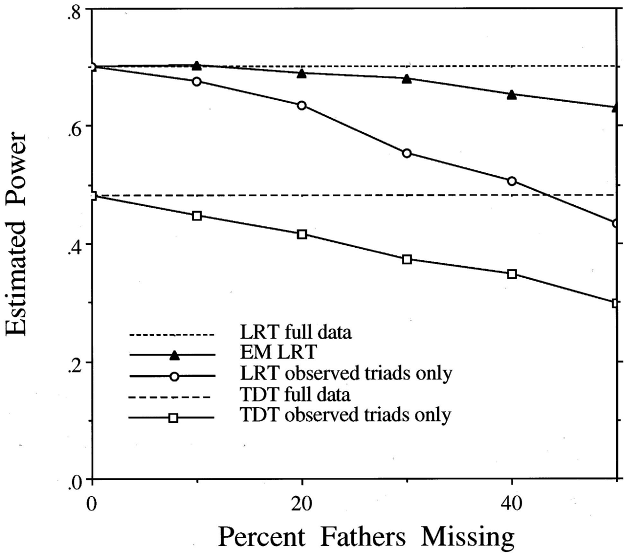

Figure 1 Simulation-based estimated powers for various testing procedures, as a function of the percentage of fathers with missing data. The points plotted indicate the empirical proportion of tests (by use of a .05 level of significance) that would have rejected the null hypothesis that $\beta_1 = \beta_2 = 0$ (equivalently, $R_1 = R_2 = 1.0$), among 1,000 simulated studies, each of which included 100 families. In the simulation of these data sets, the inherited genotype was assumed to influence risk via a recessive model (in which $R_1 = 1$ and $R_2 = 3$). The dashed lines correspond to the full-data analyses using the 2-df LRT or the TDT. The triangles plot the power resulting from application of the EM algorithm to include information from the incomplete triads.

randomly by a Bernoulli mechanism (by use of uniformly distributed random numbers), with a probability, in increments of .10 within the range .00–.50. The proportion missing was assumed to be ⩽.50, primarily because, for only 100 families, the type I error rate is >.05 when a higher proportion of fathers is missing.

For comparison, under scenarios in which $R_1 > 1$ or $R_2 > 1$, each simulated study was analyzed with full data, under both the LRT and the TDT. The analyses then were repeated, this time without the case-parent triads in which the father was flagged as missing and with only the completely observed triads. Then, the EM algorithm was applied, as described above, to recapture information from the partially observed triads, that is, the dyads comprising only the mother and the child.

Simulations were performed under the null hypothesis of no linkage disequilibrium ($R_1 = 1 = R_2$), to verify that the empirical size of a level .05 test is consistent with the nominal level. Then, data including triads with missing fathers were simulated under a dominant model ($R_1 = R_2 = 3$), a recessive model ($R_1 = 1$ and $R_2 = 3$), and a gene-dose model ($R_1 = 2$ and $R_2 = 4$). Simulations also were performed under a dominant model ($S_1 = S_2 = 3$) and a recessive model ($S_1 = 1$ and $S_2 = 3$), for a maternally mediated effect.

To demonstrate the application of the method when mothers also are missing, a set of simulations was performed with various rates of missingness for both mothers and fathers. For simplicity, the simulations were performed under the assumption that the mother and father were equally and independently likely to be missing, although the EM method does not require this assumption.

Finally, simulations were performed for families in which no fathers contributed information—that is, the information was restricted to mother/proband dyads. For these simulations, the number of families was increased from 100 to 200, because a larger number of families is required when a study is designed with mothers only, to ensure that the type I error rate is consistent with the nominal level .05.

Results of Simulations

Simulation under the null model yielded rejection rates for both the full-data LRT and the incomplete-data EM-LRT that were consistent with the .05 level. However, see Appendix B for a minor correction that is helpful, especially when a large proportion of triads are incomplete.

Figure 1 displays results for the recessive model, under a scenario in which the gene effect is inherited. As in an earlier report (Weinberg et al. 1998), the LRT (power of .7) outperformed the TDT (power <.5) when the triads were complete. With an increasing proportion of missing fathers, the power of both the LRT and the TDT declined sharply when only the completely observed triads were used in the analysis. When incomplete triads were included by means of the EM-LRT, however, the power stayed high, evidently recapturing most of the information. Figure 2 shows corresponding results under a dominant model. Again, the EM recaptured most of the lost power, with remarkably little loss even when half the fathers were missing, compared with the full-data analysis.

The TDT corresponds to the score statistic for the gene-dose model ($R_2 = R_1^2$; Schaid and Sommer 1994) and outperformed the 2-df LRT, under this scenario. Nevertheless, as shown in figure 3, the power of the TDT dropped rapidly when fathers were missing and fell below that of the EM-LRT when ⩾30% were missing. Again, the EM recaptured most of the power that would have been sacrificed by exclusion of the triads with missing data.

Results for genetic mechanisms that operate via the mother revealed more modest gains in power, for the EM algorithm (the TDT is not shown, because the TDT

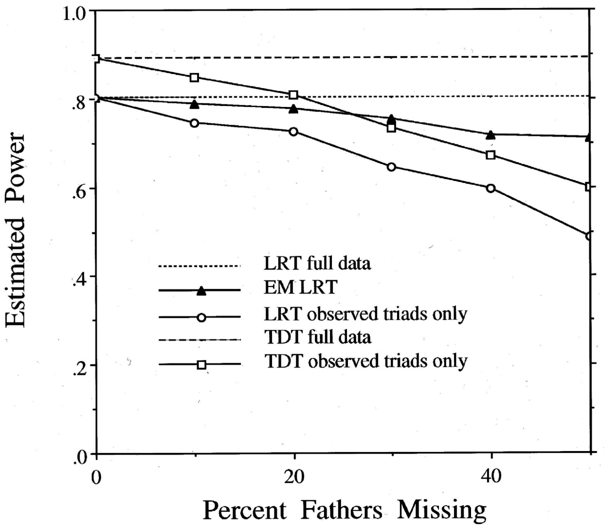

Figure 3 Plot of simulation-based estimated powers similar to those shown in figs. 1 and 2. In this set of simulations, the inherited genotype was assumed to influence risk via a gene-dose model ($R_1 = 2$ and $R_2 = 4$).

has no power beyond the type I error rate for detection of such effects). Figure 4 shows results for a recessive model ($S_1 = 1$ and $S_2 = 3$). Slightly more than half the lost power was regained by incorporation of information from the dyads, via the EM approach.

Results for a maternal-effect dominant model ($S_1 = 3$ and $S_2 = 3$) revealed somewhat better recovery of power (fig. 5). The power was excellent when the model was based on 100 triads, even with half the fathers missing.

In practice, some mothers also may be missing, and the next set of simulations was performed under the assumption that a random fraction of mothers was missing. For simplicity, mothers and fathers were assumed to be equally likely to be missing, and these events were simulated as independent. For example, in the extreme for these simulations, when the proportion of missing mothers or fathers was set to .5, an average of only 25% of triads were complete (.25 = .5 × .5), and another 25% were missing both parents. Results are shown in figure 6. Much of the lost information again was recaptured, even when mothers as well as fathers could be missing.

When all fathers were missing and the study included only cases and their mothers, the empirical level of a 2-df test was .058, on the basis of 1,000 simulations of 200 dyads, and this level is statistically consistent with the nominal level .05. The estimated power under a recessive alternative ($R_1 = 1$ and $R_2 = 3$) was .74 (compared with .96 with full data), and the estimated power under a dominant alternative ($R_1 = 3$ and $R_2 = 3$) was .97 (compared with 1.00 with full data). For a scenario in which the effect is due to the maternal genotype, the parameter S_1 can be shown to be statistically unidentifiable, because its estimate cannot be separated from that of certain other stratum parameters. Nevertheless, the LRT, which now must be based on a 1-df χ^2 distribution, remains valid. Under the null hypothesis, with 200 dy-

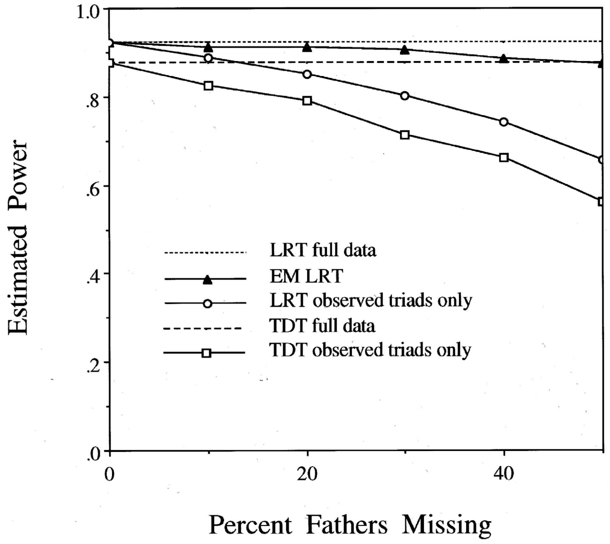

Figure 2 Plot of simulation-based estimated powers similar to that shown in fig. 1. In this set of simulations, the inherited genotype was assumed to influence risk via a dominant model ($R_1 = R_2 = 3$).

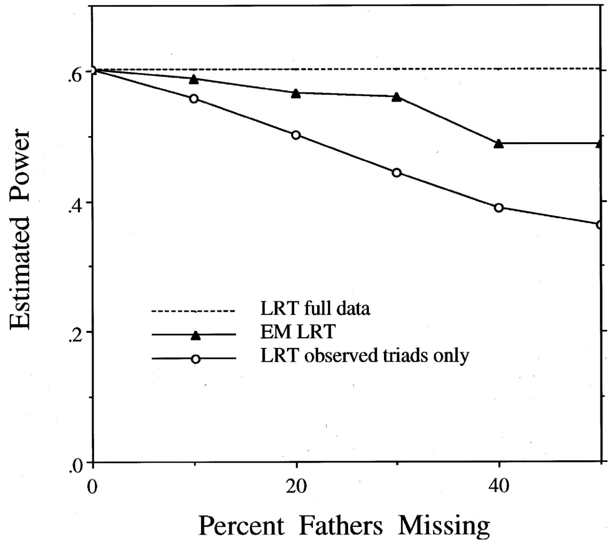

Figure 4 Plot of simulation-based estimated powers similar to those shown in figs. 1–3. In this set of simulations, the maternal genotype was assumed to influence risk via a recessive model ($S_1 = 1$ and $S_2 = 3$).

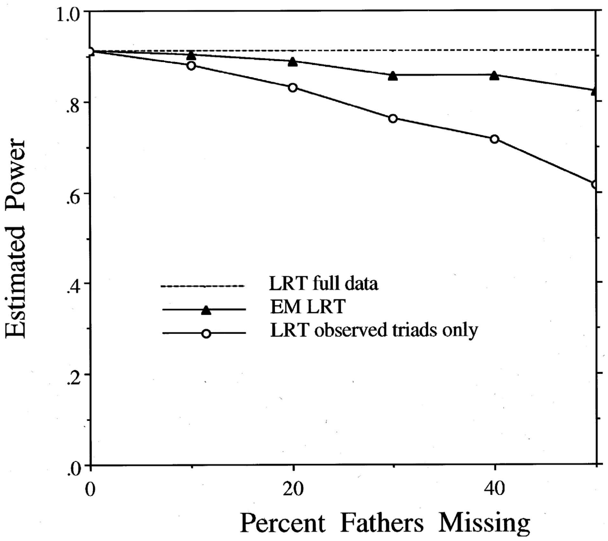

Figure 5 Plot of simulation-based estimated powers similar to that shown in fig. 4. In this set of simulations, the model was assumed to be dominant ($S_1 = 3 = S_2$).

ads, the empirical level of the test was .052, which is statistically consistent with the nominal level .05. Under a recessive alternative ($S_1 = 1$ and $S_2 = 3$) the empirical power was .52 (compared with .88 with full data). Under a dominant alternative ($S_1 = 3$ and $S_2 = 3$) the empirical power was .86 (compared with 1.00 with full data).

Discussion

When affected individuals are studied together with their parents, to test for linkage between an allelic variant or marker and an associated disease, missing parents can cause considerable loss of statistical power if the TDT is used, because incomplete triads must be discarded. In contrast, likelihood methods together with the EM algorithm allow the recovery of much of the lost information and make statistically efficient use of the data provided by dyads and monads. Simulations have revealed that, conditional on the parental genotypes, when the inherited genotype influences risk the efficiency of the EM-LRT is close to that of the full-data analysis, even when as many as half the fathers are missing. When the causal pathway is via maternal factors, the recovery of power is still substantial, although less impressive. The method works equally well when both fathers and mothers are missing, although the programming of the EM algorithm becomes more complicated.

Potentially, genetic data could be missing for some of the affected probands as well, either owing to accidents in the laboratory or, in the case of a fetus with a defect, owing to elective termination of the pregnancy and technical difficulties in obtaining usable fetal tissue or in obtaining parental consent to use the tissue. Nonetheless, in the absence of data from the affected proband, genetic data from one or two of the parents could be incorporated, via the EM algorithm, in a straightforward extension of the method described above.

If genetic data are available from siblings of the probands of some of the incomplete triads, sibling information potentially could be incorporated in the E step of the EM algorithm, thereby improving the power even more. This approach either would not collect or would not make use of the affected status of the siblings studied. If the missing individual was the father, however, one would need to verify genetically (or be willing to assume) that paternity was in fact shared between the proband and the other siblings providing genetic data.

If a multiallelic gene has been grouped into two categories for analysis, the finer allelic categories can still be used in the E step, to enable the investigator to recapture even more of the genetic information for the missing parent, thereby improving the statistical power of the LRT. Then, the M step can still be based on the biallelic grouping, under the assumption that the distinct alleles within each of the two groups share a common relationship to risk.

A technical point is that I assumed, for simplicity, that missingness is unrelated to the allele under study; how-

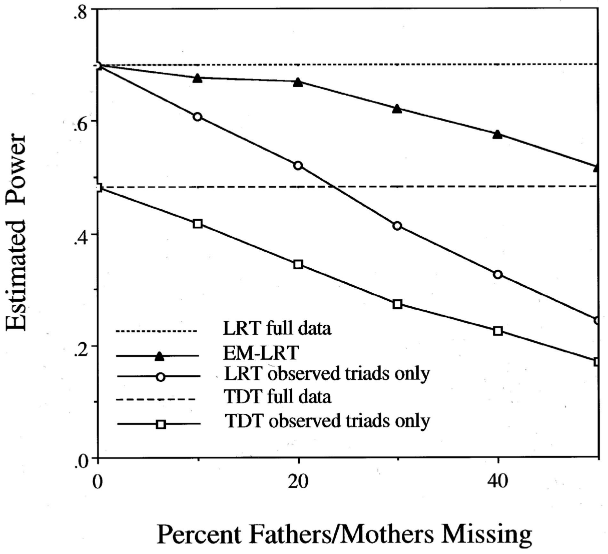

Figure 6 Plot of simulation-based estimated powers similar to that shown in fig. 1. In this set of simulations, either the mother or the father (or both) were missing. The percentage missing corresponds to a single missing probability that was applied independently to the mothers and the fathers, for simulation of the incomplete triads. For example, if the missing rates are 20%, an average of 16% of families will have data from the mother but not the father, 16% will have data from the father but not the mother, 64% will have data from both parents, and 4% will not have data from either parent.

ever, this may not always be true. Nevertheless, parameters corresponding to mating-type–specific probabilities that the father is missing could be incorporated (and estimated), and EM-based maximization using the resulting likelihood could be performed (this model would allow a test of the hypothesis that the six rates of missingness are equal, via a 5-df χ^2 test). However, if different mating-type–specific rates of missingness were required for the father and the mother, the corresponding model would be overparameterized and intractable; thus, simplification of assumptions sometimes is required. For example, the problem becomes statistically identifiable if a single rate of missingness for mothers is applied across all six mating-type categories. An alternative that might work well for some populations is to stratify on self-defined ethnicity, under the assumption of possibly different rates of missingness across the strata. Of course, there is always the option to use the analysis of complete triads, either via the TDT or the LRT, since neither method is invalidated by possibly differential rates of missing data across the mating types.

Sun et al. (1998) recently discussed the remarkable fact that linkage can be studied by use of only one parent per proband. The single-parent design can have practical advantages, because the mother may be easier to recruit than the father, and, in general, questions about parental status would not be of concern. Although Sun et al. (1998) described noniterative approaches to estimation, maximum-likelihood methods such as the EM-LRT would be expected to provide optimal efficiency. My simulations confirmed that a design based on mother/offspring dyads, analysis of which would not be possible by use of the TDT, provides good statistical power when analysis is based on the EM-LRT and when risk depends on the inherited genotype. A 1-df LRT can be applied to test for maternally mediated effects.

The single-parent design, however, imposes noteworthy limitations on inference related to linkage. Genetic mating symmetry cannot be verified on the basis of data from dyads, and the full model (eq. [1]) cannot be simultaneously fit to distinguish between effects of the maternal genotype and effects of the inherited genotype. Under a maternal-effects scenario the model is not fully identifiable, and, although a 1-df LRT is valid, parameter estimation cannot be trusted to be without bias.

Finally, because my focus in this article has been on hypothesis testing, I have not distinguished between the study of a candidate gene and the study of a marker gene. Because both scenarios reduce to the same null hypothesis, in which all the relative risks are 1.0, the EM-LRT provides a valid hypothesis test under either scenario. Thus, the distinction between a marker and a candidate gene is not important for hypothesis testing. However, if the gene under study is a candidate gene, then the model given by equation (1) can be considered as a literal representation of a multiplicative alternative to the null hypothesis. In such a context, the use of the EM algorithm will serve not only to provide a more powerful statistical test but also to improve the precision of estimation for the relative risk parameters R_1, R_2, S_1, and S_2. By contrast, if the gene under study is a marker that is related to risk of the disease only through linkage disequilibrium with a nearby risk-conferring gene, then the alternative model in table 1 is not quite correct, because the relative risk in the offspring will also depend on the parental genotype and not only on the inherited genotype, owing to recombination. Nevertheless, the EM-LRT still provides a valid test of linkage disequilibrium, relative to the null model specified by $R_1 = R_2 = 1$.

Acknowledgments

I wish to thank Drs. Norman Kaplan, Laura Scott, Marcy Speer, David Umbach, Richard Weinberg, and Allen Wilcox for their helpful comments.

Appendix A

For simplicity of exposition, I assume that only fathers are missing, but the method described can be easily generalized to accommodate missing mothers (and missing probands). Let N_{ijk} denote the observed number of triads in which the mother, the father, and the child carry i, j, and k copies, respectively, of the variant allele. Let $M_{i?k}$ denote the number(s) of dyads in which the mother has i copies, the child has k copies, and the father could not be studied and therefore carries an unknown number of copies. Let N denote the total number of families studied. Let p_{ijk} denote the cell probability for cell $(M, F, C) = (i, j, k)$—that is, the expected count of table 1 divided by N. If complete data were available, the logarithm of the multinomial likelihood would be

$$\log(L) = \sum_{\text{possible } i,j,k} N_{ijk} \log(p_{ijk}) \ . \tag{A1}$$

If some fathers are missing, the logarithm of the observed-data likelihood would be instead

$$\log(L) = \sum_{\text{possible } i,j,k} N_{ijk} \log(p_{ijk}) + \sum_{\text{possible } i,k} M_{i?k} \log\Big(\sum_{\text{possible } j} p_{ijk}\Big) \ . \tag{A2}$$

The EM algorithm provides a computational trick that allows maximization of the likelihood given in equation (A2), over the parameters of a given model—for example, the model in table 1. The general strategy calls for estimating the data and then maximizing the expression of (A1) on the basis of the estimated data. The maximization can be done by use of standard software. Let Y_{ijk}^{r} denote the estimate for the count for cell (*i, j, k*), at iteration *r,* and let p_{ijk}^{r} denote the estimate for the probability for cell (*i, j, k*), at iteration *r.* Then, the E step performs the next "estimation" as follows:

$$Y_{ijk}^{r+1} = N_{ijk} + M_{i?k} \frac{p_{ijk}^{r}}{\sum_{\text{possible } j} p_{ijk}^{r}} \ .$$

For the M step, Y_{ijk}^{r+1} is used in place of N_{ijk} in equation (A1), and the next set of p_{ijk}^{r+1} is obtained by maximization.

These two steps are repeated in their turn until the parameter estimates converge. The theory guarantees (Dempster et al. 1977) that, at each stage of iteration, the likelihood given in equation (A2)—that is, the observed-data likelihood—will increase. Following convergence to its maximum, the log likelihood then is calculated on the basis of equation (A2). When a model containing only the mating-type–stratum parameters is compared with one that also includes parameters corresponding to R_1 and R_2, the change in twice the maximized log likelihood provides a test statistic with an approximately (large samples) 2-df χ^2 distribution (under the null hypothesis), which then permits an LRT for linkage disequilibrium. Other nested models also can be compared, by contrasting, for example, the fits (via a 1-df χ^2 test) for a model that allows separately for R_1 and R_2, compared with the more constrained dominance model specifying that $R_1 = R_2$ or with a recessive model specifying that $R_1 = 1$.

Appendix B

The EM algorithm can be shown to have the property that, at each iteration, the estimated likelihood function increases. However, in certain applications in which probabilities can be estimated to be 0, the iterations can become stuck at 0 for the component of the parameter vector with a current estimate of 0. For example, if there are no observed triads in which all three members are homozygous for the variant allele, then the fitted values for mating type 1 will begin at 0 and can never move away from 0. To avoid this problem, the iterations can be bounced away from this boundary, by checking for fitted values <.01, substituting .01 in place of such extremely low values, and resuming the iteration.

References

Baker RJ, Nelder JA (1978) The GLIM system, release 3. Numerical Algorithms Group, Oxford

Curtis D, Sham PC (1995) A note on the application of the transmission disequilibrium test when a parent is missing. Am J Hum Genet 56:811–812

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B 39:1–22

Ewens WJ, Spielman RS (1995) The transmission/disequilibrium test: history, subdivision, and admixture. Am J Hum Genet 57:455–464

Falk C, Rubinstein P (1987) Haplotype relative risks: an easy, reliable way to construct a proper control sample for risk calculations. Ann Hum Genet 51:227–233

Schaid DJ, Sommer SS (1993) Genotype relative risks: methods for design and analysis of candidate-gene association studies. Am J Hum Genet 53:1114–1126

——— (1994) Comparison of statistics for candidate-gene association studies using cases and parents. Am J Hum Genet 55:402–409

Self S, Longton G, Kopecky K, Liang KY (1991) On estimating HLA-disease association with application to a study of aplastic anemia. Biometrics 47:53–61

Spielman RS, Ewens WJ (1996) The TDT and other family-based tests for linkage disequilibrium and association. Am J Hum Genet 59:983–989

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506–516

Sun F, Flanders W, Yang Q, Khoury M (1998) A new method for estimating the risk ratio in studies using case-parental control design. Am J Epidemiol 148:902–909

Weinberg CR, Wilcox AJ, Lie RT (1998) A log-linear approach to case-parent–triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. Am J Hum Genet 62: 969–978

Wilcox AJ, Weinberg CR, Lie RT (1998) Distinguishing the effects of maternal and offspring genes through studies of "case-parent triads." Am J Epidemiol 148:893–901