A New Method to Account for Missing Data in Case-Parent Triad Studies

T.L. Bergemann^a Z. Huang^b

^aBiostatistics, University of Minnesota, Minneapolis, Minn., and ^bAbbott Laboratories, Abbott Park, Ill., USA

Key Words

Log-linear models · Case-parent triad design · Missing data

Abstract

Background/Aims: The case-parent triad design is commonly used in genetic association studies. Generally, samples are drawn from an affected offspring, manifesting a phenotype of interest, as well as from the parents. The trio genotypes may be analyzed using a variety of available methods, but we focus on log-linear models because they test for genetic association and additionally estimate the relative risks of transmission. The models need to be modified to adjust for missing genotypes. Furthermore, instability in the parameter estimates can arise when certain kinds of genotype combinations do not appear in the dataset. **Methods:** In this paper, we kill two birds with one stone. We propose a new method to simultaneously account for missing genotype data and genotype combinations with zero counts. This method solves a zero-inflated Poisson (ZIP) regression likelihood. The maximum likelihood estimates yield relative risks and the information matrix gives appropriate variance estimates for inference. A likelihood ratio test determines the significance of genetic association. **Results:** We compared the ZIP regression to previously proposed methods in both simulation studies and in a dataset that investigates the risk of orofacial clefts. The ZIP likelihood estimates regression coefficients with less bias than other methods when the minor allele frequency is small.

Copyright © 2009 S. Karger AG, Basel

1. Introduction

The use of the case-parent triad design shows promise in genetic studies, particularly when studying childhood diseases with complex phenotypes. The design involves genotyping an affected offspring, as well as the offspring's parents [1, 2]. In the study of childhood cancer etiology, for example, the case-parent triad design may prove ideal because the parents are usually available and serve as a particularly motivated set of controls, population stratification is naturally corrected for, and an independent age-matched set of controls would be difficult to recruit for genetic studies. Further, programs such as the COG Childhood Cancer Research Network (CCRN) greatly assist in case identification [3]. Hypotheses for the case-parent triad design test for both linkage and/or association. A wide variety of tests are available, and have been extended to various numerical situations, including the TDT, the FBAT, conditional logistic regression models, and log-linear models [4–7].

Both the TDT and the FBAT have been extended to handle missing parental genotypes, quantitative traits, and interactions. Conditional logistic regression models can test for haplotype association, gene-gene interactions, gene-environment interactions, and maternal effects [6]. We use log-linear models in our institutional studies because this framework tests not only for linkage or association, but also estimates the level of association via a relative risk. The framework similarly allows for

tests of gene-gene interactions, gene-environment interactions, and maternal effects [7–9]. For a single genotype test, the likelihood of interest is

$$L(\beta) = \prod_{c=1}^n P(G_{m_c}, G_{f_c}, G_c | Y_c = 1). \quad (1)$$

where G_c is the offspring genotype, G_{m_c} and G_{f_c} are the maternal and paternal genotypes respectively at the same marker, and Y_c indicates the binary disease phenotype.

The likelihood in Equation 1 is maximized equivalently by solving the following log-linear model:

$$\log(E[N_{ijk}]) = \mu_{ij} + \beta_1 I[k = 1] + \beta_2 I[k = 2] + I[i = j = k = 1] \log 2 \quad (2)$$

where i, j, k indicate the genotype for the mother, father and child respectively. The outcomes N_{ijk} represent counts for each possible combination of triad genotypes from a single SNP, μ_{ij} indicates one of six possible mating types for the mother and father, and β_1, β_2 indicate the risk of disease associated with possessing one or two copies of genotype G . e^{β_1} represents the relative risk for a child with one copy of the marker allele, compared with a child born from the same type of mating pair with no copies of the allele. Similarly, e^{β_2} represents the relative risk for a child with two copies of the marker allele, similarly compared with a child having no copies. Finally, the model includes an offset term to account for the two-fold probability of transmitting one minor allele versus none or two minor alleles from a pair of heterozygous parents. Extensions of Equation 1 and 2 to estimate interactions are relatively straightforward.

When the minor allele frequency of the genotype under study is small, some combinations of genotypes in the nuclear families may not be observed. Denote the major allele of a marker as A and the minor allele as a . The minor allele frequency is p_a and then the genotype frequencies under Hardy-Weinberg equilibrium (HWE) are $P(G = 0) = (1 - p_a)^2$, $P(G = 1) = 2p_a(1 - p_a)$, and $P(G = 2) = p_a^2$. For these frequencies, the probability of observing a family where all three members have the genotype a/a is $P(N_{222}) = p_a^4$. If $p_a < 0.1$, then most reasonably sized datasets will not contain any families of this type, that is, $N_{222} = 0$. When there are zero counts of this kind in the dataset, parameter estimates in the log-linear model will become unstable. For example, in Equation 2, we see $E[N_{222}] = e^{\mu_{22} + \beta_2}$. Then, as $E[N_{222}] \rightarrow 0$, the mating type parameter becomes unstable, $\mu_{22} \rightarrow -\infty$, and the relative risk e^{β_2} will also be underestimated. Zero-inflated Poisson (ZIP) regression allows for cells with zero counts and stabilizes parameters in the regression model [10]. This

paper will tailor the ZIP likelihood to the case-parent triad design thereby stabilizing the mating type parameters.

In family-based studies, such as the case-parent triad design, the presence of missing parents, especially paternal genotypes, is a common occurrence. There are several contributing causes. For adult onset diseases, the parents of the offspring may no longer be alive. Given that husbands are usually older than wives and have shorter lifespans, the paternal genotypes are missing more often. For childhood diseases, missing data may result from paternal abandonment before disease diagnosis. A life-threatening disease diagnosis in a child may also cause familial disruption, sometimes leading to divorce, that can result in unavailable paternal information [11]. In addition, genotyping failure, an unavoidable technical issue, results in missing data.

Attempts to recover partial information in triad studies can lead to inflated type I error rates [12]. Previous studies have shown that the multiple imputation procedure, for example, has inflated type I error when 30% or more of the families have missing genotypes [13]. Thus, this paper will stress the importance of preserving the type I error rate so as to avoid inference that is too liberal. Likelihood-based methods may be employed to infer missing data. Partial information from mothers and fathers can be exploited fully. Previous likelihood-based methods have elegantly demonstrated inference in the presence of missing data while simultaneously estimating odds ratios [14]. The odds ratio approximates the relative risk when disease prevalence is low. The methods introduced here, however, will simultaneously account for missing data and estimate relative risks precisely, with no approximation required.

The expectation-maximization (EM) algorithm has been used to infer missing genotypes and recover much of the power that would be lost by eliminating families [8]. In the case-parent triad 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. Unless the EM algorithm incorporates the relative risk parameters in the missing data imputation, the method will not fully account for the variance of the relative risks. Details are given in Section 2.1. Although one can use the Louis method to correct the information matrix from the EM algorithm results, a speedier approach employs a quasi-Newton optimization instead [15]. The new methods proposed in this paper si-

multaneously estimate all important parameters, while also accounting for zero counts in the Poisson model. Section 2.2 explains a likelihood-based method that properly controls the type I error rate in the presence of missing data using the context of the log-linear model in Equation 2. The method is then expanded to the ZIP regression described in Section 2.3 to properly account for bias due to zero counts.

2. Methods

This section first outlines the EM algorithm to impute missing data for the case-parent triad design, as described nicely by Weinberg [8]. Then we provide an alternative method to optimize the same likelihood, demonstrating the ability to simultaneously account for missing data and estimate regression coefficients. Finally, this likelihood is expanded to stabilize parameters in the instance of zero counts using a ZIP model.

2.1 The EM Algorithm

Let N_{ijk} denote the observed number of complete triads, with i, j , and k denoting the number of the minor allele in the mother, father, and child, respectively. Let M_{ik} denote the number of incomplete triads, where one parent is missing, the other parent has i copies and the child has k copies of the minor allele of interest. Without loss of generality, we assume for simplicity of description, that only fathers are missing, just as in [8]. The methods are extended to scenarios with missing offspring genotypes in a straight-forward manner. Let p_{ijk} denote the probability of the occurrence of N_{ijk} among all the possible combinations. The set G contains the 15 possible genotype combinations in the triad: {000, 010, 011, 100, 101, 110, 111, 112, 021, 201, 121, 122, 211, 212, 222}. In the presence of missing data, the logarithm of the observed-data multinomial likelihood would be

$$\log(L) = \sum_{ijk \in G} N_{ijk} \log(p_{ijk}) + \sum_{ik \in G} M_{ik} \log\left(\sum_{j \in G_{ik}} p_{ijk}\right). \quad (3)$$

Then the EM algorithm is applied to impute missing data. To start with, the method fractionally assigns p_{ijk} to each possible cell, then the E-step performs the estimation as follows:

$$N_{ijk}^{r+1} = N_{ijk}^r + M_{ik} \frac{p_{ijk}^r}{\sum_{j \in G_{ik}} p_{ijk}^r}$$

where r denotes the iteration, with $r = 0, 1, 2, \dots$, until convergence. On the basis of the revised estimates of N_{ijk} , the algorithm maximizes the multinomial likelihood using the pseudocomplete data through the M step:

$$\log(L) = \sum_{ijk \in G} N_{ijk}^{r+1} \log(p_{ijk}).$$

The E step and M step repeat alternatively until convergence.

After the EM algorithm converges, the final estimated version of N_{ijk} serves as the observed N_{ijk} and is plugged into the log-linear model in Equation 2. Parameter estimates are obtained by maximizing the Poisson likelihood for the imputed

counts and calculating the relative risks. Note that this method does not account for the uncertainty arising from the imputation of family counts N_{ijk} therefore underestimating the variance of β_1 and β_2 . Instead, inference is based on a likelihood ratio test using the formulation in Equation 3 to ensure an appropriate type I error rate.

2.2. Poisson Regression Likelihood with Missing Data

The method in Section 2.1 has the advantage of stable imputation of missing data using the EM algorithm. However, it underestimates the variance of the regression coefficients because it substitutes the unobserved response with its estimates from the EM algorithm. To take this variance into account, this section provides a method that solves the full Poisson likelihood directly by incorporating both the log-linear model and the missing data uncertainty into one equation.

Recall, the log-linear model provided in Equation 2 of Section 1. After rewriting the formula, we can easily obtain the expected number of complete triads, with i maternal copies, j paternal copies and k copies from the offspring, as follows:

$$E[N_{ijk}] = e^{\mu_{ij} + \beta_1 I[k=1] + \beta_2 I[k=2] + I[i=j=k=1]} \log 2 \quad (4)$$

For complete data, the original Poisson log-likelihood is

$$\log(L_{\text{complete}} | \beta, \mu) = \sum_{ijk \in G} N_{ijk} \log E[N_{ijk}] - E[N_{ijk}] \quad (5)$$

This likelihood is equivalent to the multinomial likelihood in most situations after the Multinomial-Poisson transformation [16]. The transformation is inappropriate when parameter estimates lie on the boundary of the parameter space. This may occur when there are family counts of zero, $N_{ijk} = 0$ for some i, j, k leading to instability in the Poisson likelihood.

Rather than estimate the expected value of N_{ijk} , as in Section 2.1, we simply substitute the $E[N_{ijk}]$ with their assumed structure as in Equation 4. The Poisson likelihood is transformed into the following nice expression:

$$\begin{aligned} \log(L_{\text{complete}} | \beta, \mu) &= \sum_{ijk \in G} N_{ijk} \log E[N_{ijk}] - E[N_{ijk}] \\ &= \sum_{ijk \in G} N_{ijk} (\mu_{ij} + \beta_1 I[k=1] + \beta_2 I[k=2] + I[i=j=k=1]) \log 2 \\ &\quad - e^{\mu_{ij} + \beta_1 I[k=1] + \beta_2 I[k=2] + I[i=j=k=1]} \log 2. \end{aligned}$$

For incomplete data, the Poisson likelihood can be derived similarly.

$$\log(L_{\text{incomplete}} | \beta, \mu) = \sum_{ik \in G} \left(M_{ik} \log \sum_{j \in G_{ik}} E[N_{ijk}] - \sum_{j \in G_{ik}} E[N_{ijk}] \right). \quad (6)$$

In this way, we achieve the expression of the true likelihood without sacrificing any precision since we have not involved any estimation steps yet. Then the problem of making the observed data, including both the complete and incomplete triads, most probable is equivalent to maximizing the combined log-likelihood $\log(L_{\text{complete}} | \beta, \mu) + \log(L_{\text{incomplete}} | \beta, \mu)$ with respect to the β and μ vectors using a quasi-Newton optimization algorithm [17, 18].

Fig. 1. Comparison of Type I error rates for the three methods discussed in Section 2. The upper panels show results for $N = 300$ and the lower panels for $N = 500$. The panels from left to right show results for the EM algorithm, the direct likelihood, and the ZIP likelihood respectively.

2.3. Zero-Inflated Poisson Likelihood

As mentioned previously, when the minor allele frequency is small $p_a < 0.1$, the probability of observing zero counts in the data set is much higher. The family type N_{222} is observed in the population with frequency p_a^4 and therefore N_{222} would be unobserved with frequency $1 - p_a^4$. Assuming Hardy-Weinberg equilibrium, the frequency p_a can be used to calculate the probability of observing each family type $P[N_{ijk} | ijk \in \mathbf{G}]$ in the population. The probability of observing a zero count is then $1 - P[N_{ijk}]$ for each transmission type $\{ijk \in \mathbf{G}\}$.

Zero-inflated Poisson (ZIP) regression allows for cells with zero counts by including the probability of observing a zero count in the likelihood [10]. The general form of the ZIP regression is a finite mixture model where the responses are independent and assume a distribution as follows:

$$\begin{aligned} N_{ijk} &\sim 0 && \text{with probability } \pi_{ijk} \\ N_{ijk} &\sim \text{Poisson}(\lambda_{ijk}) && \text{with probability } 1 - \pi_{ijk} \end{aligned}$$

Under this model, when $\lambda_{ijk} = E[N_{ijk}]$, the probabilities for the count of each family type are computed as:

$$\begin{aligned} P[N_{ijk} = 0] &= \pi_{ijk} + (1 - \pi_{ijk})e^{-\lambda_{ijk}} \\ P[N_{ijk} = c] &= (1 - \pi_{ijk})e^{-\lambda_{ijk}} \lambda_{ijk}^c / c! \end{aligned}$$

We can improve upon the general ZIP model by using information about Mendelian inheritance in the likelihood. The optimization procedures are most stable when the mixing parameters map to a single value. The probabilities π_{ijk} can be treated as functions of p_a and derived assuming Hardy-Weinberg equilibrium. For example, $\pi_{000} = 1 - (1 - p_a)^4$, $\pi_{010} = 1 - p_a(1 - p_a)^3$, $\pi_{111} = 1 - 2p_a^2(1 - p_a)^2$, $\pi_{122} = 1 - p_a^3(1 - p_a)$, and $\pi_{222} = 1 - p_a^4$.

For complete data, the ZIP log-likelihood is

$$\begin{aligned} \log(L_{\text{complete}} | \beta, \mu) &= \\ \sum_{ijk \in \mathbf{G}} \log &\left(\pi_{ijk} I(N_{ijk} = 0) + (1 - \pi_{ijk}) e^{-\lambda_{ijk}} \lambda_{ijk}^{N_{ijk}} / N_{ijk}! \right) \end{aligned} \quad (7)$$

Fig. 2. Comparison of power levels for the three methods discussed in Section 2. The upper panels show results for $N = 300$ and the lower panels for $N = 500$. The panels from left to right show results for the EM algorithm, the direct likelihood, and the ZIP likelihood respectively.

where λ_{ijk} is given in Equation 4 of Section 2.2. The incomplete log-likelihood is exactly the same as shown above in Section 2.2 Equation 6. Again, a quasi-Newton algorithm is used to optimize the log-likelihood and solve for the vectors $\vec{\mu}$ and $\vec{\beta}$.

2.4. Simulation Methods

For 300 and 500 simulated family trios, we compared the performance of (1) the EM algorithm, (2) the direct Poisson likelihood, and (3) the ZIP likelihood. Performance was also assessed for the multiple imputation approach of Croiseau et al., but the results of this method only appear in the Supplemental material [13]. For each simulation, a marker is generated such that the minor allele frequency is in the set $p_a \in \{0.05, 0.10, 0.25, 0.40\}$. The proportion of missingness was simulated to occur either not at all, or in 10, 20, 30, or 40% of families in the dataset. A missing marker was assigned at random to the fathers in the simulated trios. Results will be equivalent if the missing marker is assigned to the mothers. Relative risks were set either to $e^{\beta_1} = 1$, $e^{\beta_2} = 1$ or $e^{\beta_1} = 1.5$, $e^{\beta_2} = 2.25$.

For every possible combination of the relative risks, allele frequencies, and missingness, we compare the results from 1,000 simulated datasets. In each simulation a likelihood ratio test compares the full model when $\beta_1 \neq 0$ or $\beta_2 \neq 0$ to the reduced model $\beta_1 = \beta_2 = 0$. This hypothesis tests for any association due to the genetic marker without assuming a form for penetrance. The threshold of significance in the test is set at the usual p value of 0.05.

3. Results

3.1 Simulations

For simulations where the relative risk parameters are $e^{\beta_1} = 1$, $e^{\beta_2} = 1$, one can calculate the overall type I error. The type I error should be 0.05 for each of the three methods in Section 2. Figure 1 compares the type I error

Table 1. Average relative risk estimates from simulations, $N = 300$, $e^{\beta_1} = 1.5$, $e^{\beta_2} = 2.25$

		EM algorithm		Direct likelihood		ZIP likelihood	
		$RR_1 = 1.5$	$RR_2 = 2.25$	$RR_1 = 1.5$	$RR_2 = 2.25$	$RR_1 = 1.5$	$RR_2 = 2.25$
$p_a = 0.05$							
% Missing	0	1.516	0.161	1.516	0.533	1.498	1.090
	10	1.517	0.173	1.507	0.744	1.494	1.026
	20	1.516	0.517	1.508	1.134	1.494	1.640
	30	1.525	1.231	1.519	1.574	1.503	2.268
	40	1.542	0.970	1.538	1.151	1.516	1.951
$p_a = 0.10$							
% Missing	0	1.497	2.133	1.497	2.155	1.487	2.061
	10	1.514	2.130	1.510	2.101	1.504	2.066
	20	1.503	2.103	1.501	2.113	1.492	2.050
	30	1.511	2.265	1.511	2.188	1.501	2.067
	40	1.524	2.484	1.524	2.304	1.508	2.112
$p_a = 0.25$							
% Missing	0	1.508	2.257	1.508	2.257	1.508	2.244
	10	1.517	2.307	1.518	2.288	1.515	2.280
	20	1.518	2.269	1.518	2.270	1.515	2.259
	30	1.516	2.245	1.516	2.241	1.514	2.221
	40	1.497	2.256	1.498	2.255	1.494	2.211
$p_a = 0.40$							
% Missing	0	1.493	2.219	1.493	2.219	1.492	2.218
	10	1.511	2.263	1.510	2.258	1.510	2.258
	20	1.502	2.222	1.503	2.223	1.502	2.221
	30	1.513	2.291	1.514	2.289	1.513	2.287
	40	1.510	2.285	1.511	2.286	1.512	2.285

results for each of these three methods in all scenarios. The EM algorithm given in Section 2.1 and the direct likelihood given in Section 2.2 control the type I error reasonably for all minor allele frequencies and the degree of missingness. The ZIP likelihood also controls the type I error appropriately when $N = 300$ and $p_a \geq 0.25$ or when $N = 500$ and $p_a \geq 0.1$. When the minor allele frequency is smaller than this, however, the type I error is too low. On average, when $p_a \leq 0.1$, the ZIP likelihood type I error is 0.029 when $N = 300$ and 0.037 when $N = 500$.

Next, for simulations where the relative risk parameters are $e^{\beta_1} = 1.5$, $e^{\beta_2} = 2.25$, we calculated the overall power. The highest power is expected to occur when there is no missing data and then power decreases somewhat as the amount of missing data increases. Figure 2 compares the power for all three methods in all simulation scenarios. When there is complete data, the EM algorithm and the direct likelihood yield the exact same power. The ZIP likelihood has slightly lower power than the other two

methods in this scenario. On average, with complete data, the ZIP likelihood power is 0.036 smaller than the other two methods when $N = 300$ and 0.023 smaller when $N = 500$.

As the degree of missing data increases, power will decrease. The EM algorithm shows a decrease in power between complete data and 40% missing of 0.053 when $N = 300$ and 0.035 when $N = 500$, averaged over all allele frequencies. For the direct likelihood these values are 0.056 and 0.038, respectively. The ZIP likelihood power decreases slightly more than the direct likelihood, with average decreases of 0.055 and 0.045, respectively.

Tables 1–4 in the Supplemental material (see www.karger.com/doi/000228924) provide the specific type I error and power values for each simulation scenario and method. These tables also outline the results of the multiple imputation algorithm. The results indicate that the multiple imputation procedures have less power and smaller type I error than the EM algorithm or the direct likelihood. Interestingly, this is the opposite effect than

Table 2. Average relative risk estimates from simulations, $N = 500$, $e^{\beta_1} = 1.5$, $e^{\beta_2} = 2.25$

		EM algorithm		Direct likelihood		ZIP likelihood	
		$RR_1 = 1.5$	$RR_2 = 2.25$	$RR_1 = 1.5$	$RR_2 = 2.25$	$RR_1 = 1.5$	$RR_2 = 2.25$
$p_a = 0.05$							
% Missing	0	1.501	0.722	1.501	1.228	1.487	1.478
	10	1.524	0.783	1.520	1.421	1.514	1.600
	20	1.513	1.545	1.508	1.988	1.499	2.298
	30	1.506	1.289	1.504	1.770	1.492	2.077
	40	1.507	1.584	1.506	1.814	1.488	2.224
$p_a = 0.10$							
% Missing	0	1.506	2.191	1.506	2.191	1.503	2.127
	10	1.496	2.261	1.495	2.211	1.491	2.165
	20	1.505	2.281	1.506	2.222	1.503	2.161
	30	1.500	2.318	1.500	2.244	1.497	2.145
	40	1.515	2.343	1.515	2.294	1.507	2.130
$p_a = 0.25$							
% Missing	0	1.500	2.238	1.500	2.238	1.500	2.238
	10	1.491	2.241	1.491	2.243	1.491	2.243
	20	1.499	2.247	1.499	2.244	1.500	2.246
	30	1.514	2.257	1.514	2.256	1.516	2.260
	40	1.510	2.262	1.510	2.261	1.513	2.259
$p_a = 0.40$							
% Missing	0	1.509	2.249	1.509	2.249	1.508	2.248
	10	1.495	2.253	1.495	2.253	1.495	2.252
	20	1.508	2.270	1.507	2.269	1.508	2.269
	30	1.507	2.265	1.508	2.264	1.508	2.264
	40	1.515	2.266	1.515	2.264	1.517	2.266

expected based on previous simulation studies [13]. In most situations, this algorithm also has less power and smaller type I error than the ZIP model.

Tables 1 and 2 show the average relative risk estimates in the log-additive model over 1000 simulations for each simulation scenario and for each method. Additionally Tables 5–6 in the Supplemental material show average relative risk estimates in the null model. All three optimization methods estimate the relative risk e^{β_1} fairly well, meaning that the bias of this estimate is small. The estimation of e^{β_2} , however, is much less accurate. Recall the argument in the Introduction that showed a bias when there are zero counts in the dataset. We expect the EM approach and the direct likelihood to underestimate e^{β_2} when the minor allele frequency is small and indeed this is the case. The ZIP method also underestimates e^{β_2} , but to a lesser degree. Of the three methods, when $p_a = 0.05$, the EM algorithm has the most bias and the ZIP likelihood has the least bias. Under the null model, this is true for $p_a \leq 0.10$.

Table 3 focuses on the relative risk estimate $e^{\beta_2} = 2.25$ when $N = 500$. This table provides the confidence intervals in addition to the mean estimate. The 95% confidence intervals are the simulation confidence intervals. For most parameter combinations, these intervals are the largest for the EM algorithm and the smallest for the ZIP likelihood. This provides further evidence that the ZIP likelihood tends to stabilize the relative risk estimates.

3.2. Orofacial Cleft Study

Each of the three methods in Section 2 were used to analyze a set of Danish triads where the offspring exhibited orofacial clefts. A sample of 529 families with affected offspring were genotyped for 23 single nucleotide polymorphisms (SNPs) and each was tested for association with cleft lip and palate [19]. In Shi et al. [19] the EM algorithm was used to impute missing data in the cohort. A 2 degree of freedom likelihood ratio test assessed the statistical significance of each SNP, the same test of sig-

Table 3. Average relative risk estimate for $e^{\beta_2} = 2.25$ and simulation confidence intervals, $N = 500$

		EM algorithm		Direct likelihood		ZIP likelihood	
		RR_2	95% CI	RR_2	95% CI	RR_2	95% CI
$p_a = 0.05$							
% Missing	0	0.72	$(3 \cdot 10^{-9}, 13.17)$	1.23	$(1 \cdot 10^{-5}, 13.14)$	1.48	$(1 \cdot 10^{-3}, 13.53)$
	10	0.78	$(4 \cdot 10^{-9}, 15.29)$	1.42	$(2 \cdot 10^{-4}, 13.50)$	1.60	$(1 \cdot 10^{-3}, 10.81)$
	20	1.55	$(3 \cdot 10^{-9}, 20 \cdot 10^8)$	1.99	$(2 \cdot 10^{-4}, 76554)$	2.30	$(2 \cdot 10^{-3}, 78298)$
	30	1.29	$(2 \cdot 10^{-9}, 24 \cdot 10^8)$	1.77	$(2 \cdot 10^{-4}, 51004)$	2.08	$(2 \cdot 10^{-3}, 50939)$
	40	1.58	$(2 \cdot 10^{-9}, 40 \cdot 10^8)$	1.81	$(9 \cdot 10^{-5}, 80955)$	2.22	$(2 \cdot 10^{-3}, 10 \cdot 10^5)$
$p_a = 0.10$							
% Missing	0	2.19	(0.84, 5.01)	2.19	(0.84, 5.01)	2.13	(0.88, 4.88)
	10	2.26	(0.87, 5.21)	2.21	(0.84, 5.10)	2.16	(0.87, 5.09)
	20	2.28	(0.83, 5.77)	2.22	(0.78, 5.56)	2.16	(0.80, 5.39)
	30	2.32	(0.87, 6.25)	2.24	(0.86, 5.65)	2.14	(0.85, 5.34)
	40	2.34	(0.87, 6.59)	2.29	(0.84, 6.48)	2.13	(0.83, 5.64)
$p_a = 0.25$							
% Missing	0	2.24	(1.43, 3.50)	2.24	(1.43, 3.50)	2.24	(1.42, 3.53)
	10	2.24	(1.46, 3.41)	2.24	(1.46, 3.40)	2.24	(1.47, 3.40)
	20	2.25	(1.40, 3.58)	2.24	(1.40, 3.56)	2.25	(1.41, 3.55)
	30	2.26	(1.45, 3.51)	2.26	(1.46, 3.54)	2.26	(1.46, 3.55)
	40	2.26	(1.45, 3.63)	2.26	(1.44, 3.62)	2.26	(1.46, 3.67)
$p_a = 0.40$							
% Missing	0	2.25	(1.57, 3.25)	2.25	(1.57, 3.25)	2.25	(1.56, 3.24)
	10	2.25	(1.56, 3.27)	2.25	(1.56, 3.25)	2.25	(1.56, 3.25)
	20	2.27	(1.60, 3.28)	2.27	(1.59, 3.28)	2.27	(1.59, 3.28)
	30	2.26	(1.59, 3.38)	2.26	(1.60, 3.37)	2.26	(1.59, 3.37)
	40	2.27	(1.56, 3.36)	2.26	(1.56, 3.35)	2.27	(1.56, 3.36)

Table 4. Tests of association between orofacial clefts and SNPs in the CYP1A1 ($\hat{p}_a = 0.09$), GSTA4 ($\hat{p}_a = 0.20$), and NAT2 ($\hat{p}_a = 0.30$) genes

Gene SNP	EM algorithm			Direct likelihood			ZIP likelihood		
	RR_1	RR_2	2df LRT	RR_1	RR_2	2df LRT	RR_1	RR_2	2df LRT
CYP1A1 ($N = 429$, $M = 88$)	0.637	0.972	0.028	0.632	0.866	0.027	0.624	1.021	0.029
GSTA4_snp2 ($N = 454$, $M = 69$)	0.729	0.929	0.051	0.732	0.924	0.051	0.733	0.891	0.066
NAT2_snp2 ($N = 444$, $M = 76$)	0.801	0.579	0.045	0.806	0.579	0.045	0.806	0.579	0.048

The variable N gives the number of complete triads and M gives the number of triads with a missing parental genotype.

nificance used in our simulations. These tests yielded three candidate SNPs from three different genes with promising results in the Danish triads, referred to as CYP1A1, GSTA4_snp2, and NAT2_snp2. The estimated minor allele frequency for each SNP is $\hat{p}_a = 0.09, 0.20$, and 0.30 , respectively. Here, table 4 shows test results for these SNPs using the direct Poisson likelihood and the

ZIP likelihood approaches, in addition to the previously reported EM algorithm. The table gives unadjusted p values, i.e. tests are not adjusted for multiple comparisons.

Table 4 demonstrates that all three methods give very similar results for the three SNPs tested. The sample size is between 517 and 523 families and the percent missing

ranges from 13 to 17%. Our simulations indicated that all three methods would yield similar results for this sample size and degree of missingness. The results in table 4 show that the p values are the largest using the ZIP likelihood, as expected. Further, the largest parameter estimate differences between the three methods are in the second relative risk, again as expected. Based on the results in Section 3.1, the ZIP estimates of $RR_2 = e^{\beta_2}$ are likely to be closest to the truth when the minor allele frequency is small.

4. Discussion

This research presents two new methods to test for association in case-parent triad studies. These methods simultaneously estimate relative risks, account for missing data, and test for preferential transmission of alleles from unaffected parents to affected offspring. Previous methods in the literature do not optimize this joint likelihood and therefore do not sufficiently address the variability in the data. Optimizing a joint likelihood is required to appropriately estimate the parameters and their variance. The EM algorithm shown in Section 3.1 bypasses this problem by using the multinomial likelihood for inference. Solving the direct Poisson likelihood instead gives simultaneous estimation of parameters of interest and their variance while maintaining the appropriate type I error. An added benefit is that the direct likelihood formulation may be more easily expanded to a broader class of models, including gene-gene and gene-environment interaction models.

Section 3.1 illustrated that the ZIP likelihood stabilizes relative risk estimates when the minor allele frequency is small, $p_a = 0.05$, particularly the relative risk to compare transmission of two alleles to transmission of zero alleles. Recall that the log-linear model given in Equation 2 shows that $E[N_{222}] = e^{\mu_{22} + \beta_2}$. As $E[N_{222}] \rightarrow 0$ there is increased instability in the parameter estimates μ_{22} and β_2 . The ZIP regression stabilizes these estimates and this is validated in the results in Section 3.1. This stabilization requires estimating one extra parameter in the likelihood, the minor allele frequency p_a .

Estimating an additional parameter in the ZIP likelihood results in some loss of power, anywhere between 0 and 12%, compared to the direct Poisson likelihood. Therefore, improving the accuracy of the relative risk estimates comes at the expense of power. Our recommendation is to suggest the ZIP model when bias is the primary concern in an association study and to suggest the

direct Poisson likelihood when inference is the primary concern. The power of the ZIP method, however, is still better than an approach that would remove triads with missing genotypes from the dataset. Our simulations demonstrated power gains of 10 to 18% for the ZIP method compared to methods that would ignore missing data. The direct likelihood had similar power gains between 10 and 18% compared to an approach ignoring missing data (results not shown).

The ZIP model likelihood assumes HWE when calculating the probability of zero counts for a particular family type. This assumption plays a minimal role in estimation of the key parameters, the relative risks, and in statistical inference about these parameters. Therefore, deviations from HWE, or more general mis-specification of the mixing parameters, will not noticeably affect the model or its results.

To make our methods more accessible to the statistical community, we developed a group of functions in R (<http://www.r-project.org>) to analyze case-parent triad data and infer missing data using the methods outlined in this paper. The functions are available at the website <http://www.biostat.umn.edu/~tracyb/trios.html> and are compatible with any version of R. The ZIP likelihood optimization is somewhat sensitive to the choice of initial parameters because the likelihood space has local maxima. We recommend trying several initialization vectors to ensure the validity of results. The functions available on our website also provide smart choices for the initial vector of parameters.

As these research methods develop, we are also interested in extending our approach to more general scenarios. In addition to models with interactions, we are interested in the joint estimation of haplotypes and their association with outcome in case-parent triads. Other researchers have described haplotype association methods for the case-parent design, including Guo CY et al. (2008), Allen and Satten (2007), and Gjessing and Lie (2006) [20–22]. This research provides a suitable background with which to extend haplotype inference within the ZIP likelihood approach.

In the current research environment, where scientists place great emphasis on replicating results from genetic association tests, we stress the importance of using methods that properly control type I error rates. Our research provides two new methods to test for association that do not over-estimate the rejection probabilities under the null in the presence of missing data. Studies of complex disease phenotypes also require an estimate of the impact for each covariate tested. These estimates

indicate the importance of various genetic and environmental risk factors that contribute to a complex trait. By accounting for unobserved family types in a triad dataset via the ZIP likelihood, we more accurately estimate the contribution of genetic risk factors.

Acknowledgements

The authors would like to thank our anonymous reviewers, as well as Cavan Reilly, Logan Spector, and Saonli Basu of the University of Minnesota for their insight and advice on the analysis of case-parent triad data. Thanks also to Clarice Weinberg and Min Shi of the NIEHS for providing SAS macros and the Danish orofacial cleft data. This research was supported by NIH grant 1-U01-CA122371-01 and Minnesota Medical Foundation grant 3656-9227-06.

References

- 1 Ahsan H, Hodge SE, Heiman GA, Begg MD, Susser ES: Relative risk for genetic associations: the case-parent triad as a variant of case-cohort design. *Int J Epidemiol* 2002;31:669–678.
- 2 Laird NM, Lange C: Family-based methods for linkage and association analysis. *Adv Genet* 2008;60:219–252.
- 3 Steele JR, Wellemeyer AS, Hansen MJ, Reaman GH, Ross JA: Childhood cancer research network: a North American pediatric cancer registry. *Cancer Epidemiol Biomarkers Prevention* 2006;15:1241–1242.
- 4 Spielman RS, McGinnis RE, Ewens WJ: Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 1993;52:506–516.
- 5 Horvath S, Xu X, Laird NM: The family based association test method: strategies for studying general genotype-phenotype associations. *Eur J Hum Genet* 2001;9:301–306.
- 6 Cordell HJ, Barratt BJ, Clayton DG: Case/pseudocontrol analysis in genetic association studies: A unified framework for detection of genotype and haplotype associations, gene-gene and gene-environment interactions, and parent-of-origin effects. *Genet Epidemiol* 2004;26:167–185.
- 7 Weinberg CR, Wilcox AJ, Lie RT: 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* 1998;62:969–978.
- 8 Weinberg CR: Allowing for missing parents in genetic studies of case-parent triads. *Am J Hum Genet* 1998;64:1186–1193.
- 9 Umbach DM, Weinberg CR: The use of case-parent triads to study joint effects of genotype and exposure. *Am J Hum Genet* 2000;66:251–261.
- 10 Lambert D: Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 1992;34:1–14.
- 11 Chang PN: Psychosocial needs of long-term childhood cancer survivors: a review of literature. *Pediatrician* 1991;18:20–24.
- 12 Guo CY, Cupples LA, Yang Q: Testing informative missingness in genetic studies using case-parent triads. *Eur J Hum Genet* 2008;16:992–1001.
- 13 Croiseau P, Genin E, Cordell HJ: Dealing with missing data in family-based association studies: A multiple imputation approach. *Hum Hered* 2007;63:229–238.
- 14 Dudbridge F: Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. *Hum Hered* 2008;66:87–98.
- 15 McLachlan GJ, Krishnan T: *The EM Algorithm and Extensions*, ed 2. New York, John Wiley & Sons, 2008.
- 16 Baker SG: The multinomial – Poisson transformation. *Statistician* 1994;43:495–504.
- 17 Broyden CG: The convergence of a class of double-rank minimization algorithms. *J Institute Mathematics Applications* 1070;6:76–90.
- 18 Fletcher R: A new approach to variable metric algorithms. *Computer J* 1970;13:317–322.
- 19 Shi M, Christensen K, Weinberg CR, Romitti P, Bathum L, Lozada A, Morris RW, Lovett M, Murray JC: Orofacial cleft risk is increased with maternal smoking and specific detoxification-gene variants. *Am J Hum Genet* 2007;80:76–90.
- 20 Allen AS, Satten GA: Statistical models for haplotype sharing in case-parent trio data. *Hum Hered* 2007;64:35–44.
- 21 Gjessing HK, Lie RT: Case-parent triads: estimating single- and double-dose effects of fetal and maternal disease gene haplotypes. *Ann Hum Genet* 2006;79:382–396.
- 22 Guo CY, Lunetta KL, Destefano AL, Cupples LA: Combined haplotype relative risk (CHRR): a general and simple genetic association test that combines trios and unrelated case-controls. *Genet Epidemiol* 2008;33:54–62.