American Journal of Epidemiology
Published by the Johns Hopkins Bloomberg School of Public Health 2008.

Vol. 168, No. 5
DOI: 10.1093/aje/kwn149
Advance Access publication July 23, 2008

Original Contribution

Making the Most of Case-Mother/Control-Mother Studies

M. Shi[1], D. M. Umbach[1], S. H. Vermeulen[2], and C. R. Weinberg[1]

[1] Biostatistics Branch, National Institute of Environmental Health Sciences, Research Triangle Park, NC.
[2] Departments of Endocrinology; Epidemiology, Biostatistics and HTA; and Human Genetics, Radboud University Nijmegen Medical Centre, Nijmegen, the Netherlands.

Received for publication January 30, 2008; accepted for publication May 8, 2008.

The prenatal environment plays an important role in many conditions, particularly those with onset early in life, such as childhood cancers and birth defects. Because both maternal and fetal genotypes can influence risk, investigators sometimes use a case-mother/control-mother design, with mother-offspring pairs as the unit of analysis, to study genetic factors. Risk models should account for both the maternal genotype and the correlated fetal genotype to avoid confounding. The usual logistic regression analysis, however, fails to fully exploit the fact that these are mothers and offspring. Consider an autosomal, diallelic locus, which could be related to disease susceptibility either directly or through linkage with a polymorphic causal locus. Three nested levels of assumptions are often natural and plausible. The first level simply assumes Mendelian inheritance. The second further assumes parental mating symmetry for the studied locus in the source population. The third additionally assumes parental allelic exchangeability. Those assumptions imply certain nonlinear constraints; the authors enforce those constraints by using Poisson regression together with the expectation-maximization algorithm. Calculations reveal that improvements in efficiency over the usual logistic analysis can be substantial, even if only the Mendelian assumption is honored. Benefits are even more marked if, as is typical, information on genotype is missing for some individuals.

case-control studies; genetics; linear models; polymorphism, single nucleotide; risk

Abbreviation: EM, expectation-maximization.

When studying the etiology of complex conditions with onset early in life, such as childhood cancers, certain psychiatric illnesses, congenital malformations, and pregnancy complications, both the maternal genome and the fetal genome may influence susceptibility, and both need to be considered. Case-parent triad designs, where one genotypes cases and both of their parents, can enable the investigator to differentiate fetal genetic effects from maternally mediated genetic effects (1–3) and can bypass the practical problems imposed by the need to recruit population controls. Triad designs also offer robustness against a potential source of bias called ''genetic population stratification,'' which may arise when the population consists of incompletely mixed subpopulations that differ both in their baseline disease risk (i.e., risk in people who do not carry the variant allele) and in the frequency of the genetic variant being studied. Such a population structure can produce confounding bias in a case-control study, but not in a triad study. Triad designs also permit assessment of parent-of-origin effects, where inheritance of a particular genetic variant can have effects on risk that differ according to which parent transmitted it to the offspring.

These advantages aside, triad designs suffer from some important limitations. First, fathers may be hard to recruit, and paternity is also inherently harder to be confident of than is maternity. A more disturbing limitation is that the

Correspondence to Dr. Clarice R. Weinberg, Biostatistics Branch, National Institute of Environmental Health Sciences, Mail Drop A3-03 101/A315, Research Triangle Park, NC 27709 (e-mail: weinber2@niehs.nih.gov).

TABLE 1. Expected frequencies of control mother-child pairs under Mendelian transmission of parental alleles*

	$C = 0$	$C = 1$	$C = 2$
$M = 0$	$\mu_{00} + (1/2)\mu_{01}$	$(1/2)\mu_{01} + \mu_{02}$	0
$M = 1$	$(1/2)\mu_{10} + (1/4)\mu_{11}$	$(1/2)[\mu_{10} + \mu_{11} + \mu_{12}]$	$(1/4)\mu_{11} + (1/2)\mu_{12}$
$M = 2$	0	$\mu_{20} + (1/2)\mu_{21}$	$\mu_{22} + (1/2)\mu_{21}$

* Note that μ_{mf} is proportional to the underlying frequency in the source population of parental pairs in which the mother carries m copies of the variant and the father carries f copies and where $\sum_m \sum_f \mu_{mf} = N_0$, the total number of control-mother pairs.

case-parent triad design does not permit estimation of main effects of exposures.

An alternative design calls for comparing randomly sampled mother-offspring pairs in which the offspring is healthy with mother-offspring pairs in which the offspring has the condition under study. We shall refer to this approach as the *case-mother/control-mother* design. We assume that the disease is rare in the population under study and that, although subpopulations might vary either in their baseline risks of disease or in their frequencies of the genetic variant, the covariance across subpopulations between the genotype frequency and baseline risk is 0 (4). In effect, we are making the usual assumption of no uncontrolled confounding; therefore, a case-control design is valid for this disease and population.

One complication of the case-mother/control-mother design (5) is that the maternal genome is a confounder for effects of the fetal genome, because of their correlation. Consequently, naïve analyses that use separate models to estimate effects of fetal genotypes and effects of maternal genotypes are vulnerable to confounding bias. One should instead fit a single model that simultaneously includes as predictors the fetal genotype and the maternal genotype. What has not been appreciated, however, is that the parent-child relationship implies certain linear relations among parameters. Our purpose in this paper is to describe those natural family-based constraints, to demonstrate a log-linear approach implemented through the expectation-maximization (EM) algorithm (6) that can honor them, and to document the power advantages they confer. We also assess the extent to which use of the family-based constraints can improve analytic efficiency/precision when some genotypes are randomly missing.

NATURAL CONSTRAINTS BASED ON FAMILY RELATIONSHIPS

Suppose, for simplicity, we are considering a diallelic single nucleotide polymorphism in an autosomal gene that could be related to disease susceptibility either causally or through linkage with a polymorphic causal locus. Let M and C denote the number of copies of the variant allele (i.e., 0, 1, or 2) carried by the mother and the child, respectively. It will not matter which allele is considered the "variant," but usually the one designated as such is the less frequent one, the "minor" allele. One obvious constraint that applies to both case pairs and control pairs is that (M,C) cannot be (2,0) or (0,2), because a homozygous mother has to pass on one of her two identical alleles to her child. Thus, instead of nine mother-child pairs being possible, only seven are possible.

Considering the father, who is not directly studied in this design, there are nine possible pairs of parental genotypes. Let μ_{mf} denote the population frequency of pairs of parents in which the mother has m copies and the father f copies of the allelic variant. Suppose control mother-child pairs are selected at random from the source population, where transmission from mother to offspring follows Mendelian inheritance and survival to the time of study is nondifferential by genotype. If the disease is rare or unrelated to the variant under study, then the population-based distribution of mother-child paired genotypes among controls can be expressed in terms of the μ_{mf} parameters and Mendelian proportions (table 1). We have simply collapsed over the missing fathers. For example, the (0,0) cell in table 1 consists of triads with (M,F,C) equal to (0,0,0) and (0,1,0) with expected frequencies μ_{00} and half of μ_{01}, respectively.

With no additional assumptions about the population, the $M = 1$ row already implies a constraint: The expected counts for (1,0) and for (1,2) sum to the expected count for (1,1). Thus, the family relationship alone specifies two structural zeroes and also a constraint.

Next, suppose that in addition to Mendelian inheritance we assume parental *mating symmetry* in the source population, at the locus under study (i.e., $\mu_{mf} = \mu_{fm}$ for all m, f). This additional assumption reduces the nine original μ_{mf} parameters in table 1 to only six. Adjusting the cell components of table 1 accordingly, the family relationships then imply a second constraint for the expected counts for mother-child pairs, (M,C): The expected difference between the count for (1,0) and the count for (0,1) equals the expected difference between the count for (1,2) and the count for (2,1)—namely, $(1/4)\mu_{11} - \mu_{02}$, which is the same as $(1/4)\mu_{11} - \mu_{20}$.

Another constraint that is often plausible is parental *allelic exchangeability*, which asserts that in the source population, conditional on the set of four alleles carried by a pair of parents, those alleles are randomly allocated to the two individuals. This condition is a single-locus special case of

TABLE 2. Expected frequencies of case mother-child pairs under a multiplicative model for risk*

	$C = 0$	$C = 1$	$C = 2$
$M = 0$	$B[\mu_{00} + (1/2)\mu_{01}]$	$BR_1[(1/2)\mu_{01} + \mu_{02}]$	0
$M = 1$	$BS_1[(1/2)\mu_{10} + (1/4)\mu_{11}]$	$(1/2)BR_1S_1[\mu_{10} + \mu_{11} + \mu_{12}]$	$BR_2S_1[(1/4)\mu_{11} + (1/2)\mu_{12}]$
$M = 2$	0	$BR_1S_2[\mu_{20} + (1/2)\mu_{21}]$	$BR_2S_2[\mu_{22} + (1/2)\mu_{21}]$

* Note that μ_{mf} denotes the underlying frequency in the source population of parental pairs in which the mother carries m copies of the variant and the father carries f copies. R_1 and R_2 denote the relative risks for a child with one or two copies, respectively, relative to a child with no copies; S_1 and S_2 denote the relative risks for a child whose mother has one or two copies, respectively, relative to the child whose mother has no copies. B is a normalizing constant included to ensure that the expected counts will sum to the total number of case-mother pairs.

parental *haplotype exchangeability* (7). This assumption is slightly stronger than mating symmetry, but it is much weaker than Hardy-Weinberg equilibrium because it permits the existence of genetically distinct subpopulations. Under parental allelic exchangeability, because there are four ways to assign one variant each to two parents, $\mu_{11} = 4\mu_{02} = 4\mu_{20}$. This exchangeability assumption also implies the other two assumptions, and it follows that the expected difference between the count for (1,0) and the count for (0,1) and the expected difference between the count for (1,2) and the count for (2,1) are not just equal to each other but are both equal to 0. Thus, with this slightly stronger additional assumption, now three constraints can be imposed on the expected counts for control pairs.

What about the distribution for case-mother pairs? Under a multiplicative model for risk of a rare condition, the expected counts for case mother-child pairs can be expressed in terms of the μ_{mf} parameters, Mendelian proportions, and relative risks (table 2). Here R_1 and R_2 are the relative risks for a child with one or two copies, respectively, relative to a child with no copies, and S_1 and S_2 are the relative risks for a child whose mother has one or two copies, respectively, relative to a child whose mother has no copies. The parameter B is the normalizing constant included to ensure that the expected counts sum to the total number of case-mother pairs.

FITTING MODELS THAT ENFORCE THESE CONSTRAINTS

In the usual logistic regression model, one conditions on all the predictor variables and models the log odds of disease. This approach is wonderfully flexible because one does not need to specify the distribution of covariates when maximizing the relevant conditional likelihood. The downside is that one has no way to impose prior knowledge about that covariate distribution. Imposing appropriate constraints on the covariate distribution can improve statistical efficiency (3, 8).

One way to impose constraints on the covariate distribution is to use log-linear Poisson regression. If no constraints are imposed, the results of fitting logistic and Poisson regression models are identical for the same data set. Using logistic regression to carry out a case-mother/control-mother analysis with the (M,C) count data corresponding to tables 1 and 2, one would fit the model:

$$\ln\left(\frac{\Pr(D \mid M,C)}{1-\Pr(D \mid M,C)}\right) = \mu + \beta_1 I_{(C=1)} + \beta_2 I_{(C=2)} + \alpha_1 I_{(M=1)} + \alpha_2 I_{(M=2)}.$$

Here, $I_{(\text{expression})}$ is an indicator function which is 1 when the expression is true and 0 when it is false. The coefficients β_1 and β_2 are the natural logarithms of R_1 and R_2, while α_1 and α_2 are the natural logarithms of S_1 and S_2.

To accomplish an equivalent analysis using Poisson regression, let N_{mcd} denote the observed number of families in which $M = m$, $C = c$, and $D = d$, where d is 1 for case pairs and 0 for control pairs, and let $E(N_{mcd})$ be the expected value of that count. One uses the 14 observed cell counts to fit the following Poisson regression model:

$$\ln[E(N_{mcd})] = \theta_{mc} + \delta d + \beta_1 dI_{(c=1)} + \beta_2 dI_{(c=2)} + \alpha_1 dI_{(m=1)} + \alpha_2 dI_{(m=2)}.$$

The seven parameters θ_{mc}, one for each (M,C) cell among controls, by allowing complete flexibility for the control-mother distribution (consider setting $d = 0$), ensure that the covariate distribution is unconstrained. An advantage to using the Poisson version of these two identical approaches is that, by modeling the cell counts directly, the Poisson approach provides a way to impose constraints on the θ_{mc} parameters describing the covariate distribution.

An additional difficulty is that the constraints we have described are linear constraints on the cell counts or, equivalently, on the μ_{mf} parameters, but they are nonlinear constraints on the θ_{mc} parameters, because those parameters are the natural logarithms of the cell counts. Imposition of such nonlinear constraints is not straightforward in available software packages like Stata or SAS. Other software, for example, LEM (log-linear expectation maximization) by van den Oord and Vermunt (9), easily handles such constraints.

For the constraints that we are considering, it is convenient to imagine an idealized data structure with 15 cells for case-parent triads (as in the article by Weinberg et al. (1)) and a similar data structure with 15 cells for control-parent

triads, but where the fathers' genotypes are all missing. One then can use the EM algorithm to maximize the fatherless likelihood, and it becomes easy to impose these constraints. The assumption required by the EM algorithm that the fathers' genotypes be noninformatively missing is trivially satisfied because all are missing. The proposed construction automatically satisfies the linear constraint (and the structural zeroes) that follows from Mendelianism and the family relationships. One must, of course, use the observed-data likelihood, rather than the pseudo-complete-data likelihood, to compute the likelihood ratio χ^2 statistic.

To impose parental mating-type symmetry, one collapses each 15-cell multinomial to a 10-cell multinomial (1), because, for example, the triple genotype (0,1,*C*) is merged with (1,0,*C*). One can then again use the EM algorithm to maximize the appropriate likelihood. The additional constraint of parental allele exchangeability (7) can be honored by using the same 10-cell multinomials but using a single stratum parameter for both the {0,2} and {1,1} parental strata and assigning offsets that are the logarithms of 2, 1, 2, and 1 to the *MFC* triads (0,2,1), (1,1,0), (1,1,1), and (1,1,2), respectively. (Here "(0,2,1)" includes "(2,0,1)," because parental switches are treated as equivalent under mating symmetry.)

In an actual case-parent/control-parent study, a proportion of the genotypes will be missing, either because the individual was not studied (e.g., the baby did not survive, or umbilical cord blood but not maternal blood was retained) or because the laboratory could not assign the genotype. Not only does the Poisson approach together with the EM algorithm (6) permit imposition of constraints, it facilitates the use of partial data when genotypes are missing. For such an approach to be valid, one must assume that missingness is noninformative—that is, missingness is random conditional on disease status and the observed genotypes. Thus, if some offspring genotypes are missing due to failure to survive, one must assume that survival is unrelated to the unobserved genotype among case mother-offspring pairs and also unrelated to the unobserved genotype among control mother-offspring pairs.

POWER COMPARISONS

To evaluate the power gains possible by exploiting various constraints in the analysis of a case-mother/control-mother study, we considered a study of 150 case-mother pairs and 150 control-mother pairs. For convenience, we employed a source population in which the single nucleotide polymorphism was in Hardy-Weinberg equilibrium. Hardy-Weinberg equilibrium is neither necessary nor assumed in our analyses, but it simplifies power calculations by allowing us to specify the μ_{mf} parameters as simple functions of allele frequency. A source population in Hardy-Weinberg equilibrium satisfies all three assumptions. Assuming the model of tables 1 and 2, we examined several risk scenarios defined by R_1, R_2, S_1, and S_2 over a range of allele frequencies.

For complete data, we calculated power for the usual logistic regression analysis (no constraints) and for our proposed analysis under each of the three nested levels of constraints. We also calculated power for a case-parent triad design with 150 cases. We repeated these calculations for scenarios with 20 percent of the genotypes randomly missing. For these missing-genotype scenarios, we considered two versions of the unconstrained logistic analysis: one restricted to mother-offspring pairs with complete genotype data and one where pairs with missing data were included via Poisson regression and the EM algorithm. Constrained analyses always included pairs with missing data.

We studied the noncentrality parameter, equivalently the power, for the 4-df χ^2 likelihood ratio test of the null hypothesis that $R_1 = R_2 = S_1 = S_1 = 1$. We made use of the fact that the noncentrality parameter of the likelihood ratio test statistic under a specified alternative can be closely approximated by the likelihood ratio statistic calculated by treating the expected counts under that alternative as if they were data (10). We calculated expected cell counts under various scenarios using the formulae in tables 1 and 2 and employed LEM software (9) to maximize the observed data likelihoods under the null and alternative hypotheses. LEM software is freely available, and the reader can download the LEM "scripts" that we used for maximizing the relevant observed data likelihoods under any of our three sets of assumptions from our website (http://www.niehs.nih.gov/research/atniehs/labs/bb/staff/weinberg/index.cfm#downloads).

We plotted the noncentrality parameters as a function of allele frequency for analyses under different sets of assumptions and included horizontal reference lines corresponding to specific power values for a 0.05-level 4-df likelihood ratio test. When the noncentrality parameter exceeds a particular cutpoint, the power exceeds the specified power. To modify the number of cases studied to some other number, say *K*, one can simply multiply these noncentrality values by *K*/150. The ratio of any two of the case-mother/control-mother curves corresponds to the relative efficiency of the two analytic approaches, across allele frequencies—that is, the approximate ratio of sample sizes required to achieve any desired level of statistical power.

Consider first a scenario in which R_1, R_2, S_1, and S_2 are 2, 3, 1, and 1, respectively. This scenario includes a gene-dose effect of the fetal genotype with no effects of the maternal genotype. Under this scenario, noncentrality curves for analyses that impose the constraints lie above the curve for the usual logistic regression analysis (figure 1, panel A). Simply imposing the fact that the mother-offspring pairs reflect Mendelian proportions improved power, particularly at allele frequencies below 0.5. Imposing two constraints or all three together improved the power even more. For this scenario, a case-parent design with 150 case triads provided power comparable to that of an unconstrained case-mother/control-mother logistic analysis with 150 case-mother pairs and 150 control-mother pairs.

We obtained qualitatively similar results with two additional risk scenarios. In a scenario where the only effect is a recessive effect of the fetal genotype (R_1, R_2, S_1, and S_2 are 1, 3, 1, and 1, respectively), increasing the number of imposed constraints again increased the power across all allele frequencies (figure 1, panel B). The power advantage of even the simple familial constraint was marked. For this

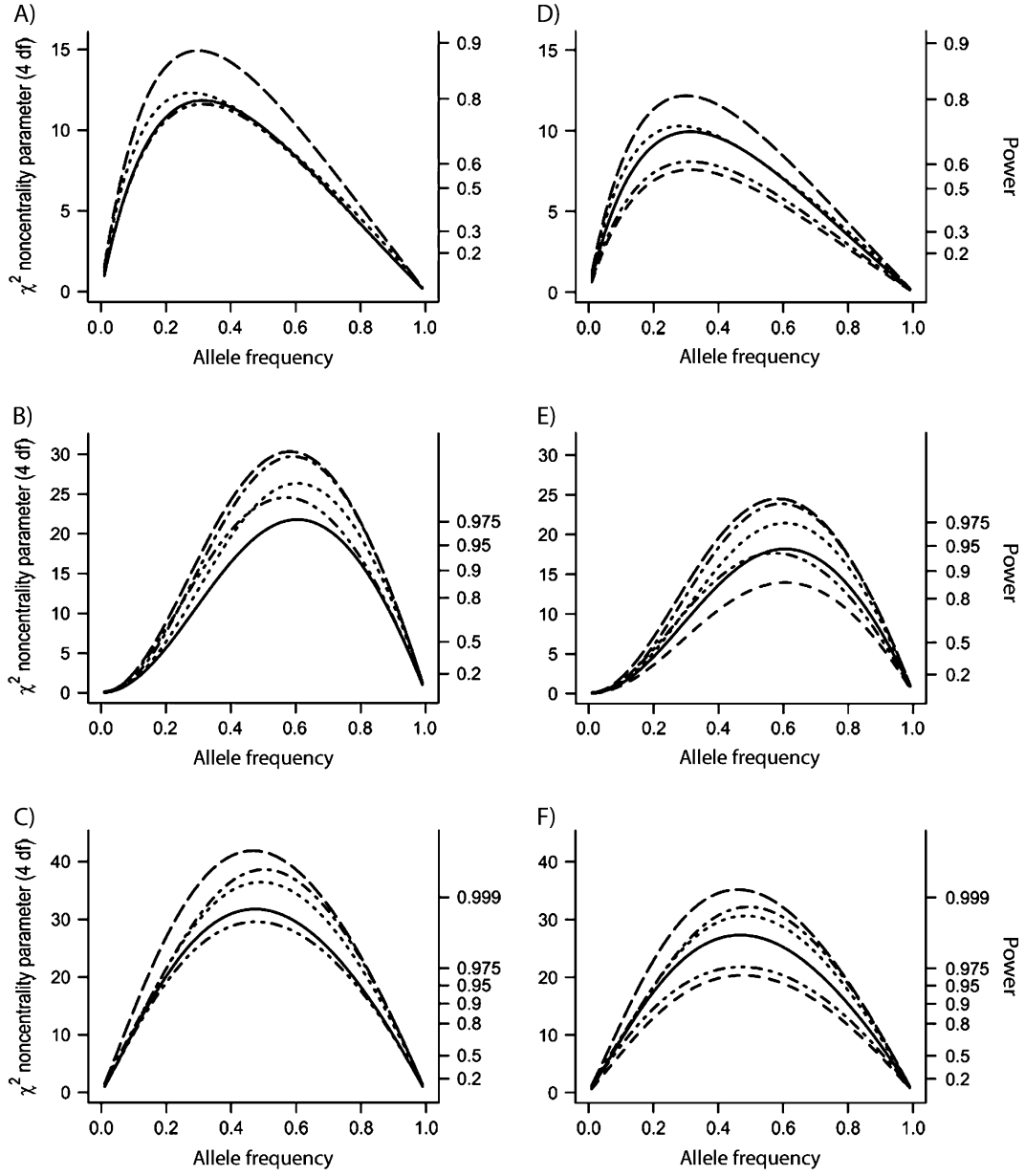

FIGURE 1. Noncentrality parameter and power as a function of allele frequency for the case-mother/control-mother design and case-parent triad design. The vertical axes show, in the left column, the χ^2 noncentrality parameter for a 4-df likelihood ratio test, and, in the right column, the power of a corresponding test with $\alpha = 0.05$. Left column (panels A–C): no missing data; right column (panels D–F): 20% of genotypes missing. First row (panels A and D): $R_1 = 2$, $R_2 = 3$, $S_1 = 1$, $S_2 = 1$; second row (panels B and E): $R_1 = 1$, $R_2 = 3$, $S_1 = 1$, $S_2 = 1$; third row (panels C and F): $R_1 = 1$, $R_2 = 3$, $S_1 = 2$, $S_2 = 2$. Curves for a case-mother/control-mother design with 150 case-mother pairs and 150 control-mother pairs: logistic regression using all pairs (solid line: —), logistic regression omitting pairs with missing genotypes (short-dashed line: – – – (panels D–F only)), log-linear Poisson regression using all pairs and imposing only the family relationship constraint (dotted line: - - -), similar analysis that additionally imposes mating symmetry (dashed-dotted line: – - –), and similar analysis that additionally imposes parental allelic exchangeability (long-dashed line: — — —). Curve for a case-parent triad design with 150 triads: log-linear Poisson regression using all triads (dashed-dotted-dotted line: – - - –). For panels A and D, curves for the model imposing mating symmetry (dashed-dotted line: – - –) and the model imposing parental allelic exchangeability (long-dashed line: — — —) overlap.

scenario, the power of the triad design exceeded that of the unconstrained case-mother/control-mother analysis and even exceeded that of some constrained analyses at low allele frequency. In a scenario where the fetal genotype has a recessive effect and the maternal genotype has a dominant effect (R_1, R_2, S_1, and S_2 are 1, 3, 2, and 2, respectively), power again increased as more constraints were imposed on the analysis (figure 1, panel C). In this scenario,

however, the case-parent triad design with 150 cases had lower power than the unconstrained case-mother/control-mother analysis. These results indicate that, when justified, imposing these plausible constraints can markedly improve analysis for case-mother/control-mother studies.

We revisited the same three risk scenarios when 20 percent of the genotypes were missing (figure 1, panels D, E, and F, respectively). Although power was predictably lower when genotypes were missing, the relations among different case-mother/control-mother analyses that used the EM algorithm to include pairs with some genotypes missing was much the same as the relations observed when no genotypes were missing. In essence, missing genotypes reduce the effective sample size by a constant fraction for those analyses, but the EM algorithm assures that the loss will be less than the full 20 percent. Also predictably, the unconstrained analysis that included pairs with some missing genotypes was more powerful than the unconstrained analysis that included only pairs with complete genotypes. With 20 percent missing genotypes but all families included in the analysis, the efficiency of the case-parents design suffered markedly. This decline may reflect the fact that, when genotypes are missing at random, the proportion of triads missing at least one genotype is larger than the proportion of mother-child pairs missing at least one genotype, leading to a greater loss of efficiency. Again, even with no additional assumptions, exploiting the family relationship markedly improves the power when some genotypes are missing.

DISCUSSION

By undertaking a case-control study, the investigator commits to some key assumptions required for valid analysis. The case-control comparison is only free of bias in the absence of genetic population structure or, in its presence, if the allele frequencies and baseline risks do not covary with baseline risks across subpopulations (4). Otherwise, bias remains a threat and one should consider the more robust case-parent triad design. Even with a triad design, however, one still must assume parental mating symmetry in order to assess maternally mediated genetic effects. In the context of a case-mother/control-mother design, one can assess maternally mediated effects without this parental symmetry assumption or, as we have shown, exploit this same assumption to improve efficiency. To take advantage of the implicit constraints this assumption imposes on parameters, one can use a log-linear Poisson regression analysis in combination with the EM algorithm instead of the usual logistic regression analysis. Parental allele exchangeability can additionally be imposed to gain even more efficiency. Surprisingly, even if one only imposes the family relationship, with no additional assumptions, the improvement in efficiency can be marked.

The power for the case-mother/control-mother design is then comparable to and often much better than that of the case-parent triad design with the same number of cases, although slightly more genotyping is required (genotyping four people instead of three people for each case if the case:control ratio is 1). The case-mother/control-mother design requires participation of control pairs, who may be hard to recruit, but it offers an additional advantage in that the main effects of covariates can also be studied.

The inclusion of covariates in the log-linear analysis is a bit trickier than it would be with logistic regression. For a categorical exposure, one can stratify, ideally allowing parental mating type parameters to be different at different levels of the exposure. For a simple dichotomy, a constrained analysis would build on the approach described previously (3) but with separate imposition of the desired constraints for each exposure-specific control-mother multinomial.

Although the assumptions we have entertained are natural and plausible and thus seem widely applicable, they might be incorrect for certain genes in some populations. If the assumptions are imposed inappropriately, estimated relative risks will be biased. Because these assumptions imply constraints on the expected counts of table 1, they can be probed by means of statistical tests. However, we have not investigated the operating characteristics of such tests.

In summary, several plausible assumptions arise from considering mother-offspring pairs in the context of nuclear families. Here we have shown how constraints derived from those assumptions can be enforced in the analysis of case-mother/control-mother studies. Whether genotype data are missing or not, use of the log-linear Poisson model enables one to enforce constraints that exploit the mother-offspring relationship in order to inform the expectation step of the EM algorithm and thereby improve the efficiency of analysis, even without any additional assumptions. If additional assumptions are adopted, the power advantages for the constrained analyses are even greater.

ACKNOWLEDGMENTS

This research was supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences. It was also supported by the Ter Meulen Fund.

The authors thank Drs. Grace Kissling, Aimee D'Aloisio, and Laura Mitchell for their helpful critiques.

Conflict of interest: none declared.

REFERENCES

1. Weinberg CR, Wilcox AJ, Lie RT. A log-linear approach to case-parent triad data: assessing effects of disease genes that act directly or through maternal effects, and may be subject to parental imprinting. Am J Hum Genet 1998;62:969–78.
2. Wilcox AJ, Weinberg CR, Lie RT. Distinguishing the effects of maternal and offspring genes through studies of "case-parent triads." Am J Epidemiol 1998;148:893–901.
3. Umbach D, Weinberg C. The use of case-parent triads to study joint effects of genotype and exposure. Am J Hum Genet 2000;66:251–61.
4. Wacholder S, Rothman N, Caparaso N. Population stratification in epidemiologic studies of common genetic variants and

cancer: quantification of bias. J Natl Cancer Inst 2000;92: 1151–8.
5. Posey DL, Khoury MJ, Mulinare J, et al. Is mutated MTHFR a risk factor for neural tube defects? Lancet 1996;347:686–7.
6. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B 1977;39:1–38.
7. Shi M, Umbach D, Weinberg C. Identification of risk-related haplotypes using multiple SNPs from nuclear families. Am J Hum Genet 2007;81:53–66.
8. Chatterjee N, Kalaylioglu Z, Carroll R. Exploiting gene-environment independence in family-based case-control studies: increased power for detecting associations, interactions and joint effects. Genet Epidemiol 2005;28:138–56.
9. van den Oord E, Vermunt J. Testing for linkage disequilibrium, maternal effects, and imprinting with (in)complete case-parent triads, by use of the computer program LEM. Am J Hum Genet 2000;66:335–8.
10. Agresti A. Categorical data analysis. New York, NY: John Wiley & Sons, Inc, 1990.