American Journal of Epidemiology
Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health 2011.

Vol. 173, No. 5
DOI: 10.1093/aje/kwq363
Advance Access publication:
February 4, 2011

Practice of Epidemiology

Using Imputed Genotypes for Relative Risk Estimation in Case-Parent Studies

Min Shi, Stephanie J. London, Grace Y. Chiu, Dana B. Hancock, Dmitri Zaykin, and Clarice R. Weinberg*

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

Initially submitted May 26, 2010; accepted for publication September 28, 2010.

Meta-analyses of genome-wide association studies are often based on imputed single nucleotide polymorphism (SNP) data, because component studies were genotyped using different platforms. One would like to include case-parent triad studies along with case-control studies in such meta-analyses. However, there are no published methods for estimating relative risks from imputed data for case-parent triad studies. The authors propose a method for estimating the relative risk for a variant SNP allele based on a log-additive model. Their simulations first confirm that the proposed method performs well with genotyped SNP data. As an empirical test of the method's behavior with imputed SNPs, the authors then apply it to chromosome 22 data from the Mexico City Childhood Asthma Study (1998–2003). For chromosome 22, the authors had data on 7,293 SNPs that were both genotyped and imputed using the software MACH, which relies on linkage disequilibrium with nearby SNPs. Correlation between estimated relative risks based on the actual genotypes and those based on the imputed genotypes was remarkably high ($r^2 = 0.95$), validating this method of relative risk estimation for the case-parent study design. This method should be useful to investigators who wish to conduct meta-analyses using imputed SNP data from both case-parent triad and case-control studies.

epidemiologic methods; genome-wide association study; genotype; imputation; meta-analysis; risk

Abbreviation: SNP, single nucleotide polymorphism.

Genome-wide association studies are widely used for mapping genes related to complex diseases, in both case-control and family-based designs. Meta-analyses can then be used to assess the consistency of results across studies, both in the identified alleles and in the direction of their estimated effects. Such analyses allow investigators to combine the overall evidence in order to quantify the relative risks associated with risk-related alleles. To efficiently pool information from studies that used different genotyping platforms with different sets of single nucleotide polymorphisms (SNPs), researchers impute genotypes at untyped SNPs by modeling the correlation structure among SNPs using a set of reference genotypes such as HapMap (1). The software package MACH (2) implements one such imputation method. Instead of the 3 genotype categories 0, 1, or 2, corresponding to the number of variant alleles present, MACH gives posterior probabilities for the 3 genotypes as well as the imputed genotype, as a numeric score in the interval [0, 2], representing a probability-weighted average of the 3 possible genotypes.

The scoring of imputed genotypes poses challenges for combining data from case-control and family-based studies. Traditionally, the analysis for case-parent data is transmission-based, employing a log-linear (3) or conditional logistic (4) regression model for estimation of relative risks. One can also use the ratio of counts of heterozygous parents who do (numerator) versus do not (denominator) transmit the variant allele (5). However, such approaches cannot be readily implemented with imputed genotypes. In this paper, we propose an alternative approach that uses imputed genotypes to estimate the relative risk for case-parent studies. We first develop the method and subsequently assess its performance with simulated genotype data. We then apply it to data from an actual case-parent triad study, allowing us to compare results based on imputed SNPs with results based on measured genotypes for the same SNPs in the same triads.

Table 1. Probabilities for Specific Case-Parent Triads Under a Log-Additive Genetic Model[a]

No. of Allele Copies			Complement	*2C-M-F*	Probability[b]	*E*(*2C-M-F*|Parents)[c]
Mother (*M*)	Father (*F*)	Child (*C*)				
2	2	2	2	0	$\mu_{22}R_1^2$	0
2	1	2	1	1	$\mu_{12}R_1^2$	$(R_1 - 1)/(R_1 + 1)$
2	1	1	2	−1	$\mu_{12}R_1$	
1	2	2	1	1	$\mu_{12}R_1^2$	
1	2	1	2	−1	$\mu_{12}R_1$	
2	0	1	1	0	$\mu_{02}R_1$	0
0	2	1	1	0	$\mu_{02}R_1$	
1	1	2	0	2	$\mu_{11}R_1^2$	$2(R_1 - 1)/(R_1 + 1)$
1	1	1	1	0	$2\mu_{11}R_1$	
1	1	0	2	−2	μ_{11}	
1	0	1	0	1	$\mu_{01}R_1$	$(R_1 - 1)/(R_1 + 1)$
1	0	0	1	−1	μ_{01}	
0	1	1	0	1	$\mu_{01}R_1$	
0	1	0	1	−1	μ_{01}	
0	0	0	0	0	μ_{00}	0

[a] R_1 represents the relative risk with 1 copy of the risk allele.

[b] μ_{ij} represents the mating type parameter for $M = i$ and $F = j$ or $M = j$ and $F = i$.

[c] Expected case-complement difference given parental genotypes.

MATERIALS AND METHODS

Scenarios with measured genotypes

First, consider scenarios in which the SNP marker is typed. Let M, F, and C represent the number of copies of the designated allele carried by the mother, father, and affected offspring in a triad, respectively. (It is of no consequence which allele is enumerated, except that interpretation of the relative risk is inverted under the alternative choice.) We define the *complement* as the hypothetical matched sibling who carries each pair of alleles that were not transmitted to the affected child. The complement would carry the genotype $M + F - C$ at each locus. Let R_1 and R_2 denote the relative risks associated with inheritance of 1 or 2 copies of the designated allele, respectively, relative to inheritance of no copies. We assume a log-additive model for risk such that $R_2 = R_1^2$. Note that under a case-parent triad design, it is relative risks that are being estimated, not odds ratios.

Let D_i represent the following adjusted case-complement difference for the ith triad: $D_i = (2C_i - M_i - F_i)I_{(M+F=1 \text{ or } 3)} + (0.5)\ (2C_i - M_i - F_i)I_{(M=F=1)}$. Here, we are using the notation I_{event}, which becomes 1 when the ‘‘event’’ statement is true and 0 otherwise. One can then take the average of these D_i's using only the triads that have at least 1 heterozygous parent, that is, the informative triads. (The triads in which both parents are homozygous are noninformative for relative risk estimation.) Under the log-additive model, it can be shown algebraically that among informative triads the expected value of D_i is $(R_1 - 1)/(R_1 + 1)$ (Table 1). Let $\bar{D}$ denote this average, which has the expected value $(R_1 - 1)/(R_1 + 1)$. Note that, following standardization of $\bar{D}$, we can square this statistic and compare it with tables for the 1-df chi-squared distribution to perform a hypothesis test for association. It also follows that under a log-additive model, we can estimate R_1 as $(1 + \bar{D})/(1 - \bar{D})$ and R_2 as the square of that. We base the standard error of $\bar{D}$ on the empirical variance of D_i. According to the central limit theorem (6), the corresponding standardized statistic is approximately normally distributed, and 95% confidence intervals for $(R_1 - 1)/(R_1 + 1)$ can be constructed. The upper and lower bounds of the 95% confidence interval for R_1 can also be calculated by back-calculating those bounds.

Scenarios with imputed genotypes

Unlike a measured genotype, an imputed genotype can be any value between 0 and 2, and the triad type is no longer confined to the 15 cells in Table 1. Another complication of imputed genotypes is that many imputation methods do not take family structure into account, and thus the imputed genotypes may violate Mendelian inheritance or be close to violation—for example, a triad with imputed genotypes $M = 1.9$, $F = 1.9$, and $C = 0.05$. To estimate the relative risk, a method needs to exclude or down-weight such implausible triads, as well as nearly noninformative triads, while taking the uncertainty of the imputations into account.

Instead of the 15 cells of Table 1, now consider the 2 × 2 × 2 cube with 27 nodes, where the axes are imputed genotypes M, F, and C (Figure 1). For a genotyped marker, M, F, and C can each take 3 different values, and the triad

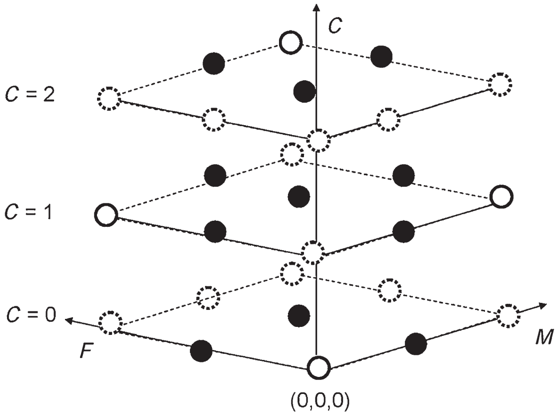


Figure 1. A 2 × 2 × 2 cube containing nodes that correspond to possible triad genotypes for a diallelic single nucleotide polymorphism (circles with solid edges), some of which are informative because there is at least 1 heterozygous parent (filled-in circles). This cube also contains nodes with Mendelian inconsistencies (open circles with dotted edges). The imputed genotypes for triads lie within or on the surface of the cube.

genotype (M, F, C) is located at a node in the cube that corresponds to one of the 15 cells of Table 1; 11 of these nodes correspond to informative triad genotypes (solid circles) and 4 to noninformative triad genotypes (open circles with solid borders). The remaining 12 nodes are impossible under Mendelian transmission (open circles with dotted borders). For an imputed marker, the triad genotypes can be located anywhere on or inside the cube. We seek to include triads clustering around the 11 informative nodes (filled circles in Figure 1) for relative risk estimation, while discounting triads that are farther away and closer to an open circle.

We achieve this differential weighting in the following way. For each triad, we first measure the minimum distance to one of the 16 non-Mendelian or noninformative nodes, denoting this distance as S. We want to keep the triads that are far away from such nodes (i.e., those with S larger than a cutoff value, say L) and down-weight those with smaller S. The weight is calculated as $W = (S/L)I_{(S<L)} + I_{(S \geq L)}$. We found empirically that a cutoff of $L = 0.75$ works well, in that it produces good correlation between the measured results and the imputed results. Triads having imputed genotypes that are closest to a non-Mendelian node are particularly subject to imputation error; therefore, we down-weight those triads even more, by multiplying S by half before calculating the weight. We then normalize the resulting weights across triads to add up to 1, by dividing each by their sum.

As noted above, the expected case-complement difference is higher by a factor of 2 for triads with 2 heterozygous parents. Therefore, we again must use a corrected D_i, obtained by multiplying the case-complement difference by the factor

$$k_i = 0.5 + 0.5\sqrt{(M_i - 1)^2 + (F_i - 1)^2}$$

For measured genotypes in informative triads, k_i is equal to the inverse of the number of heterozygous parents. For imputed genotypes, k_i takes a value between 0.5 and 1.3.

Using also the weights defined above, we now take the weighted average of the corrected differences, denoted $\bar{D}_w$. We can use $\bar{D}_w$ to estimate R_1 just as we use $\bar{D}$ with measured genotypes, using the fact that R_1 is approximately equal to $(1 + \bar{D}_w)/(1 - \bar{D}_w)$. A hypothesis test of R_1 equaling 1 can be based on testing of this average statistic against its null expectation of 0.

An alternative and less statistically defensible approach that is sometimes advocated simply treats the ''most likely'' genotype as if it were real data. After excluding evident Mendelian inconsistencies among the triads, one then estimates the relative risk using the number of transmissions of the variant allele from heterozygous parents to the number of nontransmissions from heterozygous parents.

Estimating R_1 and R_2 separately with imputed genotypes

We have assumed that the overall analysis will be carried out assuming a log-additive risk model, which will tend to enhance efficiency of testing based on a large number of SNPs. Once the meta-analysis has been completed and particular loci have been identified as related to risk for the disease under study, one may wish to return to the data for those particular SNPs and now estimate relative risks under a codominant model, rather than impose the log-additive constraint. To accomplish separate estimation of R_1 and R_2, one can use a multiple imputation procedure (7). MACH outputs the posterior probabilities of the 3 genotypes for each individual at each SNP. One can use these for each triad to sample the genotypes of the mother, father, and offspring separately based on the corresponding probabilities. This assigns each scored triad to one of the nodes in Figure 1. We can then fit the imputed data with a log-linear model with parameters R_1 and R_2 in the model. For the ith imputation, one can compute the point and variance estimates for the parameter vector $(\hat{R}_1^i, \hat{R}_2^i)$. After m such imputations have been completed and analyzed, the average of the m point estimates $(\bar{R}_1, \bar{R}_2)$ provides the overall estimate (R_1, R_2). Let $\bar{U}$ denote the average of the m within-imputation variance-covariance matrices, and let B denote the among-imputation variance-covariance matrix, which is the empirical variance among the m vector estimates $(\hat{R}_1^i, \hat{R}_2^i)$. Then the estimated variance associated with the overall estimate ($\bar{R}_1$, $\bar{R}_2$) is given by $T = \bar{U} + (1 + 1/m)B$.

Simulation study and results

We first confirmed the validity of our method by applying it to measured genotypes. We generated triad data sets as previously described (8). We randomly sampled alleles for the 2 parents according to an assigned SNP frequency. We then randomly created a child from the parents on the basis of Mendel's law and assigned a relative risk of disease to the child based on the number of inherited copies of the risk allele. We calculated the risk of the disease and assigned disease status to the offspring at random on the basis of that risk. Only families with an affected offspring were retained.

Table 2. Relative Risk Estimates Based on Measured Genotypes (0, 1, or 2) Under Simulated Scenarios With Varying Relative Risks and Allele Frequencies[a]

True R_1	Allele Frequency	R_1	95% Confidence Interval	Empirical Coverage
1	0.1	1.01	1.00, 1.02	0.951
	0.2	1.00	0.99, 1.01	0.956
	0.3	1.01	1.00, 1.01	0.959
	0.4	1.00	0.99, 1.01	0.952
	0.5	1.00	0.99, 1.00	0.954
1.2	0.1	1.20	1.19, 1.21	0.944
	0.2	1.20	1.19, 1.21	0.953
	0.3	1.20	1.19, 1.21	0.950
	0.4	1.20	1.19, 1.20	0.957
	0.5	1.20	1.20, 1.21	0.949
1.5	0.1	1.50	1.48, 1.51	0.941
	0.2	1.51	1.50, 1.52	0.956
	0.3	1.51	1.50, 1.52	0.947
	0.4	1.51	1.50, 1.52	0.951
	0.5	1.49	1.48, 1.50	0.940

[a] Each simulated data set contained 400 triads, and 1,000 data sets were simulated under each scenario.

For simulation efficiency, we imposed Hardy-Weinberg equilibrium, though our method does not require this assumption. We assumed a log-additive risk model and generated data sets with measured genotypes under 3 relative risk scenarios: $R_1 = 1$, $R_1 = 1.2$, and $R_1 = 1.5$, each with 5 different allele frequencies: 0.1, 0.2, . . ., 0.5. For each scenario, we simulated a data set with 400 families 1,000 times. We checked for bias by calculating the simulation-based overall confidence interval based on the empirical standard deviation for R_1 across simulations.

Under all simulation scenarios, our method estimated R_1 with no evident bias, and the 95% coverage rates for study-based confidence intervals were statistically consistent with the nominal 95% coverage rates (Table 2). Log-linear analysis, being maximum-likelihood-based, is optimal for measured genotypes. The confidence interval for our proposed method was 5%–6% wider, on average, than that of the log-linear modeling approach. This simulation study demonstrated the validity of our method for measured genotypes and revealed the expected slight loss of information in comparison with the full likelihood method.

Example

A genome-wide association study based on 492 Mexican children with asthma and their parents was conducted in 1998–2003 (9), using the Illumina HumanHap 550K BeadChip (Illumina, Inc., San Diego, California) to identify novel genetic variants associated with childhood asthma. Nonparentage was excluded in these 492 trios by analyses of the genome-wide association data using PLINK (10). To enable pooling of genome-wide association data from several studies, we conducted imputation to generate approximately 3 million autosomal SNP imputed genotypes. The software MACH 1.0 (http://www.sph.umich.edu/csg/abecasis/MaCH/index.html) was used to perform the imputation, with the HapMap Phase 2 release 21 consensus haplotypes used as the referent (ftp://ftp.hapmap.org/hapmap/phasing/2006-07_phaseII/consensus/). We included the CEU, YRI, JPT, and CHB HapMap populations as reference groups and did not apply a minor allele frequency filter (11). The total set of approximately 3 million SNPs targeted for analysis was based on the SNPs included in HapMap for those ethnic groups. To impute genotypes using MACH, one compiles a list of SNPs and MACH uses the measured SNPs to impute a genotype for each SNP on the list, using the local linkage disequilibrium structure, even if the SNPs were measured. We consequently had both measured and imputed genotypes for a subset of the SNPs. We used the SNPs on chromosome 22 as our example. There were 7,293 SNPs for which both imputed and measured genotypes were available, plus an additional 28,116 that were imputed. We applied our proposed approach with $L = 0.75$ to estimate R_1 using imputed genotypes and compared the estimates with the corresponding estimates obtained from applying a log-linear model (3) using the measured genotypes.

We investigated the sensitivity of the R_1 estimation to the selection of the cutoff value L by repeating the analysis on the same set of 7,293 SNPs with $L = 0.25$. We also analyzed the 22,074 chromosome 22 SNPs (restricting the data to those with a minor allele frequency greater than 0.02) for which only imputed genotypes were available, using 2 different cutoff values for L: 0.25 and 0.75.

We applied the multiple imputation approach to estimate R_1 and R_2 separately, using data on the 5,391 SNPs that had minor allele frequencies greater than 0.1 (to ensure estimability of R_2) and also had both measured and imputed genotypes. We performed 100 imputations for each SNP. For comparison, we also estimated R_1 and R_2 by means of the log-linear model, using measured genotypes directly.

RESULTS

As Figure 2 shows, estimates from the proposed method (with $L = 0.75$) using imputed genotypes and the log-linear model using measured genotypes were highly correlated ($r^2 = 0.95$). The 95% confidence intervals based on imputed genotypes contained the point estimates based on the measured genotypes for all of the SNPs and vice versa. With $L = 0.25$, the correlation between the estimates obtained using imputed genotypes and those obtained using measured genotypes remained high ($r^2 = 0.95$) (Appendix Figure 1). Results based on the ‘‘most likely’’ genotypes using the transmission-based ratio were also highly correlated with those based on the same ratio, but with the measured genotypes ($r^2 = 0.95$) (see Web Figure 1, which is posted on the *Journal*’s Web site (http://aje.oxfordjournals.org/)).

The R_1 estimates obtained with different L’s were highly correlated for the set of 7,293 SNPs with both measured and imputed genotypes ($r^2 = 0.999$), as well as for the set of 22,074 SNPs with only imputed genotypes ($r^2 = 0.95$ when using this set). Figure 3 shows the differences between the R_1 estimates

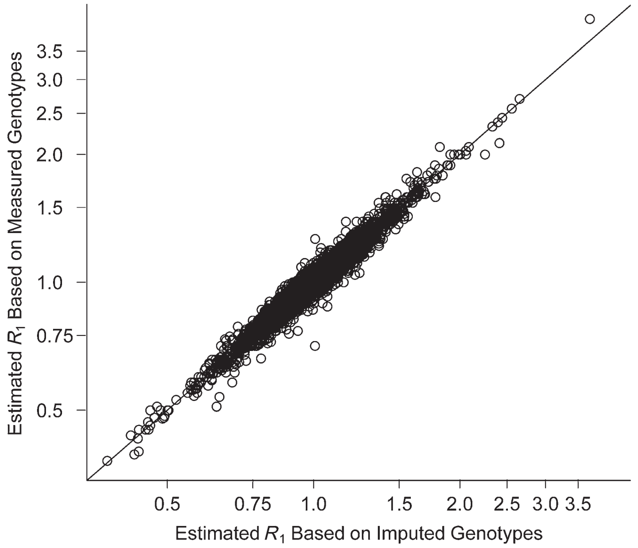


Figure 2. Estimated R_1 on a logarithmic scale based on the measured genotypes compared with that obtained using the imputed genotypes instead. The estimates were based on a log-additive model, and the cutoff value of the distance S was 0.75. There were 7,293 chromosome 22 single nucleotide polymorphisms represented with both measured and imputed genotypes from the Mexico City Asthma Study (9).

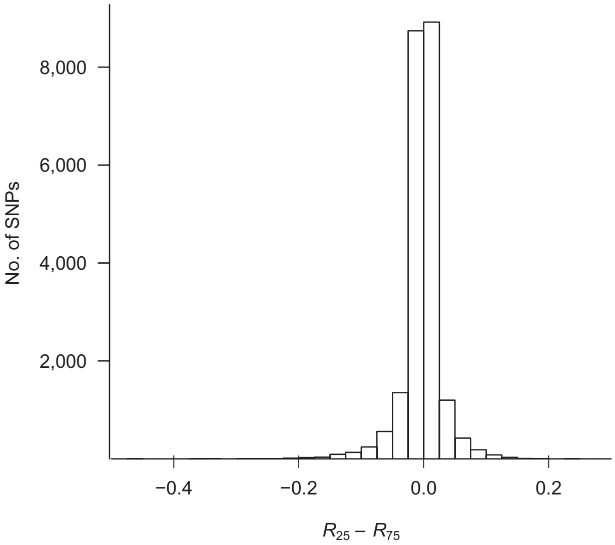


Figure 3. Difference between 2 R_1 estimates when they were analyzed with 2 different cutoff values of the distance S, $L = 0.25$ and $L = 0.75$ (see text for details). This histogram includes 22,074 single nucleotide polymorphisms (SNPs) on chromosome 22 with minor allele frequencies greater than 2% and no measured genotype. R_{25} is the R_1 estimate with $L = 0.25$, and R_{75} is the R_1 estimate with $L = 0.75$.

when they were estimated with $L = 0.25$ versus $L = 0.75$, for the SNP set with only imputed genotypes. The differences were between −0.2 and 0.2 for 99.85% of the SNPs. The 33 SNPs that showed a larger difference all had a minor allele frequency less than 0.05. Thus, the results appeared to be insensitive to the choice of the tuning parameter, L.

When we used multiple imputation to estimate R_1 and R_2 separately, 2 of the SNPs did not have enough data for estimating R_2 and were subsequently removed. As Figure 4 and Figure 5 show, estimates from the multiple imputation approach using imputed genotypes and the log-linear model using measured genotypes for the remaining 5,389 SNPs were highly correlated for both R_1 and R_2 ($r^2 = 0.98$ and $r^2 = 0.97$, respectively). These high correlations probably reflect a better fit of the model allowing for 2 relative risk parameters.

DISCUSSION

Our proposed method can be used to estimate relative risks under a log-additive model using imputed genotypes from case-parent triads. These estimates can subsequently be used for meta-analysis. This method also provides a way to increase the number of triads that can be included for a particular estimation of relative risk, by using imputation software to fill in SNPs that are sporadically missing (those that cannot be ‘‘called’’).

Perhaps the simplest available method for using imputed genotypes involves simply selecting the most likely genotype at each locus and then treating that as if it were measured genotype data. That is, one selects the highest of the 3 posterior probabilities and treats the corresponding allele count as if it were a measured genotype. In fact, for many SNPs, the linkage disequilibrium structure is strong enough that the most likely genotype corresponds to the actual genotype most of the time. For our chromosome 22 data, there was disagreement between the measured and ‘‘most likely’’ genotypes for approximately 1 in every 2,000 SNPs. The corresponding proportion of SNPs/triads with apparent Mendelian inconsistencies was 1/3,000. Thus, it is not surprising that log-linear analyses yielded high correlations for results based on our method and results based on these assigned values (excluding Mendelian inconsistencies). One issue with using ‘‘most likely’’ genotypes, however, is possible underestimation of the uncertainties associated with relative risk estimation. A related issue is that the accuracy of the ‘‘most likely’’ genotypes can vary across SNPs. For some SNPs in our chromosome 22 data, the percentage of disagreement between the measured genotype and the ‘‘most likely’’ genotype was as high as 21%, and the percentage of apparent Mendelian inconsistencies ranged up to 4.7%. For these few badly imputed SNPs, which cannot readily be identified, our method outperformed the ‘‘most likely’’ method, presumably because it effectively removes families with poor imputation genotype scores—those residing far away from the 15 nodes.

Although our results suggest that imputed data performed very well in these analyses, it is worth noting that relative risks are estimated with case-parent data, while odds ratios are estimated with case-control data. For a common disease such as asthma, this distinction could require a nullward correction of the case-control estimates (or an inflation away from the null of the triad-based estimates) to ensure full comparability before combining evidence across data from case-control and case-parent study designs. One would need

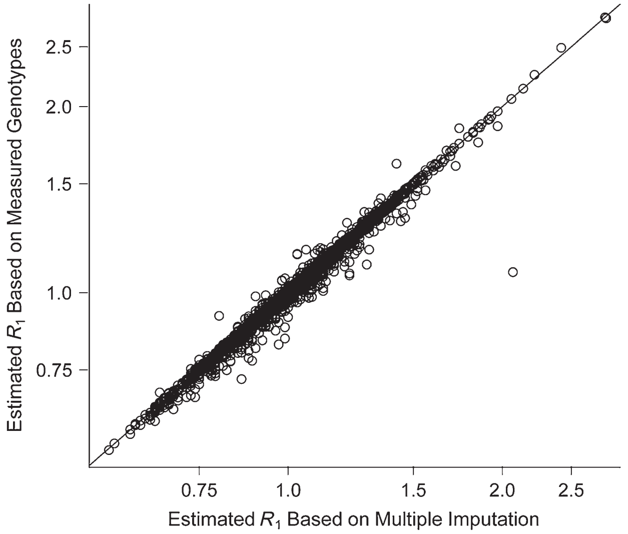


Figure 4. Estimated R_1 on a logarithmic scale based on the measured genotypes compared with that obtained using instead the imputed genotypes based on multiple imputation.

a way to estimate b_0, the baseline risk, in order to translate a relative risk to the corresponding odds ratio. For example, suppose that b is the proportion of children who develop asthma in the population under study and p is the prevalence of the putative risk-related variant allele, with a relative risk of R_1 for a single copy and ${R_1}^2$ for 2 copies. We then can apply Hardy-Weinberg equilibrium to approximate $b = (1-p)^2b_0 + 2p(1-p)b_0R_1 + p^2R_2b_0$, so the baseline risk b_0 is approximately equal to $b/[(1-p)^2 + 2p(1-p)R_1 + p^2R_2]$. Then, the odds ratio for inheritance of a single copy would be estimated by plugging the estimates for b_0 and R_1 into $R_1(1-b_0)/(1-R_1b_0)$.

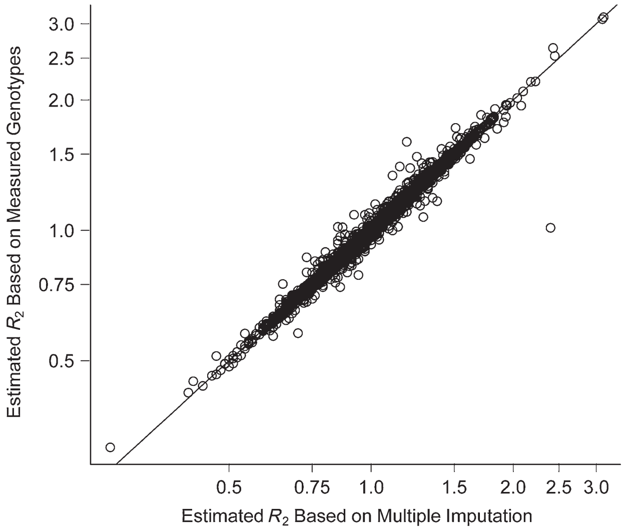


Figure 5. Estimated R_2 on a logarithmic scale based on the measured genotypes compared with that obtained using instead the imputed genotypes based on multiple imputation.

The choice of the cutoff value L determines the weight given to each triad with a non-nodal score. A high cutoff value gives less weight to triads that are not very close to any of the 11 informative nodes, while a low value allows triads relatively far from those nodes to contribute. Our sensitivity analysis suggests that results are not particularly sensitive to the choice of this parameter.

It is possible to modify our method by instead using posterior probabilities for the 3 possible genotypes directly. In this alternative approach, each family in the sample contributes a sum of increments that are weighted by the product of 3 respective genotype probabilities. For example, the ith family will contribute to the statistic an amount equal to

$$\Delta_i = \sum_{2c-m-f\neq 0} \Pr(M_i = m)\Pr(F_i = f)$$

$$\Pr(C_i = c)\left(\frac{1}{2}\right)^{I_{(m=f)}}(2c-m-f).$$

Let the total weight for the ith triad be defined by

$$W_i = \sum_{2c-m-f\neq 0} \Pr(M_i = m)\Pr(F_i = f)\Pr(C_i = c).$$

Once contributions of all triads in the sample are added, the statistic is estimated as $\bar{\Delta} = \sum_i \Delta_i / \sum_i W_i$, and the risk estimate is $(1+\bar{\Delta})/(1-\bar{\Delta})$. This approach provides risk estimates that are similar to those from the method described in Materials and Methods assuming a log-additive risk model, but its r^2 with estimates based on genotyped markers was not quite as high as that based on the method described above ($r^2 = 0.93$ vs. $r^2 = 0.95$).

The current version of the MACH program does not make inferential use of the family structure, and development of software for imputation that exploits that structure could presumably work even better for analysis of family data. The posterior probabilities from such imputed genotypes, together with multiple imputation, could also be used for mapping of genetic variants associated with quantitative traits (e.g., by employing a likelihood-based method, such as quantitative polytomous logistic regression (12)) or could be used for extended pedigrees (e.g., by employing a method that can handle data with such pedigree structures, such as the pedigree disequilibrium test (13)).

For a condition with onset in early life or for a pregnancy complication, the maternal genome may also contribute to risk, so it may be of interest to carry out meta-analyses based on possible maternal effects (14). This can also be done using the multiple imputation method we have described. If both maternal and offspring-based effects are found for the same SNP, one can fit log-linear models that include both, as well as characterize possible synergistic effects between the mother and her offspring (15).

In summary, we have shown via simulations and a real data example that estimates based on our approach applied to imputed genotypes agree well with results based on the corresponding measured genotypes using maximum likelihood estimation and a log-linear model. It is reassuring that estimation based on imputed genotypes is consistently very close to what would have been estimated on the basis of actual genotyping. This method will be useful to consortia of investigators who wish to combine data from case-control and case-parent genome-wide association studies in meta-analyses.

ACKNOWLEDGMENTS

Author affiliations: Biostatistics Branch, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina (Min Shi, Dmitri Zaykin, Clarice R. Weinberg); Epidemiology Branch, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina (Stephanie J. London, Dana B. Hancock); and Westat Inc., Research Triangle Park, North Carolina (Grace Y. Chiu).

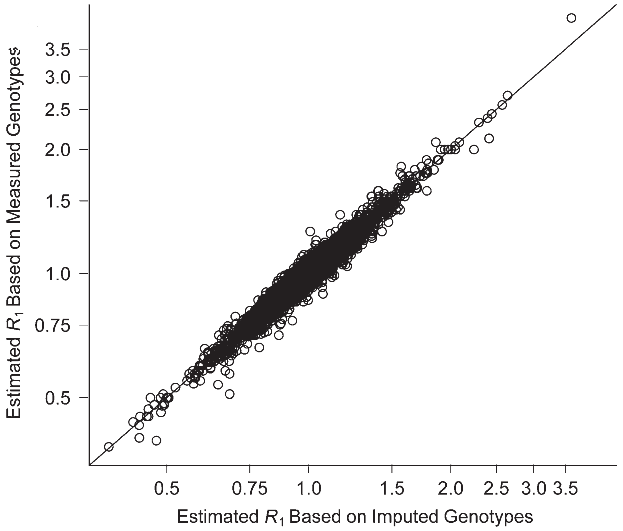
This research was supported by the Intramural Research Program of the National Institutes of Health, National Institute of Environmental Health Sciences (NIEHS), under project numbers ZO1 ES040007, ZO1 ES49109, and ZO1 ES240245.

The authors thank Drs. Isabelle Romieu, Juan-Jose Sienra Monge, Blanca del Rio Navarro, and Matiana Ramirez for their role in enrollment of study subjects, Drs. Huiling Li and Nell Burch of the NIEHS for expert technical assistance, and Joshua Smith and Dr. Debbie Nickerson of the University of Washington for performing the genome-wide association study genotyping. For data management assistance, the authors thank Dr. Shuangshuang Dai and John Grovenstein of the NIEHS. Finally, the authors thank Dr. David Umbach for his comments on the manuscript.

Conflict of interest: none declared.

REFERENCES

1. International HapMap Consortium. The International HapMap Project. *Nature*. 2003;426(6968):789–796.
2. Li Y, Willer C, Sanna S, et al. Genotype imputation. *Annu Rev Genomics Hum Genet.* 2009;10:387–406.
3. 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(4): 969–978.
4. Cordell HJ, Clayton DG. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. *Am J Hum Genet.* 2002;70(1):124–141.
5. 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(3):506–516.
6. Freedman D, Pisani R, Purves R. *Statistics*. 3rd ed. New York, NY: W W Norton & Company; 1997.
7. Rubin D. *Multiple Imputation for Nonresponse in Surveys*. New York, NY: John Wiley & Sons, Inc; 1987.
8. Shi M, Umbach DM, Weinberg CR. Identification of risk-related haplotypes with the use of multiple SNPs from nuclear families. *Am J Hum Genet.* 2007;81(1):53–66.
9. Hancock DB, Romieu I, Shi M, et al. Genome-wide association study implicates chromosome 9q21.31 as a susceptibility locus for asthma in Mexican children. *PLoS Genet.* 2009; 5(8):e1000623. (doi: 10.1371/journal.pgen.1000623).
10. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007;81(3):559–575.
11. Huang L, Li Y, Singleton AB, et al. Genotype-imputation accuracy across worldwide human populations. *Am J Hum Genet.* 2009;84(2):235–250.
12. Kistner EO, Weinberg CR. Method for using complete and incomplete trios to identify genes related to a quantitative trait. *Genet Epidemiol*. 2004;27(1):33–42.
13. Martin ER, Monks SA, Warren LL, et al. A test for linkage and association in general pedigrees: the pedigree disequilibrium test. *Am J Hum Genet.* 2000;67(1):146–154.
14. 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(9):893–901.
15. Sinsheimer JS, Palmer CG, Woodward JA. Detecting genotype combinations that increase risk for disease: maternal-fetal genotype incompatibility test. *Genet Epidemiol*. 2003; 24(1):1–13.



Appendix Figure 1. Estimated R_1 on a logarithmic scale based on the measured genotypes compared with that obtained using the imputed genotypes instead. The estimates were based on a log-additive model, and the cutoff value of the distance S was 0.25. There were 7,293 chromosome 22 single nucleotide polymorphisms represented with both measured and imputed genotypes from the Mexico City Asthma Study (9).