

2009

Multiple Imputation to Correct for Measurement Error in Admixture Estimates in Genetic Structured Association Testing

Miguel A. Padilla

Old Dominion University, mapadill@odu.edu

Jamin Divers

Laura K. Vaughan

David B. Allison

Hemant K. Tiwari

Follow this and additional works at: https://digitalcommons.odu.edu/psychology_fac_pubs



Part of the [Genetics Commons](#), and the [Statistical Models Commons](#)

Repository Citation

Padilla, Miguel A.; Divers, Jamin; Vaughan, Laura K.; Allison, David B.; and Tiwari, Hemant K., "Multiple Imputation to Correct for Measurement Error in Admixture Estimates in Genetic Structured Association Testing" (2009). *Psychology Faculty Publications*. 56. https://digitalcommons.odu.edu/psychology_fac_pubs/56

Original Publication Citation

Padilla, M. A., Divers, J., Vaughan, L. K., Allison, D. B., & Tiwari, H. K. (2009). Multiple imputation to correct for measurement error in admixture estimates in genetic structured association testing. *Human Heredity*, 68(1), 65-72. doi:10.1159/000210450

Multiple Imputation to Correct for Measurement Error in Admixture Estimates in Genetic Structured Association Testing

Miguel A. Padilla^{a,b} Jasmin Divers^c Laura K. Vaughan^d David B. Allison^{d,e}
Hemant K. Tiwari^d

Departments of ^aPsychology and ^bMathematics and Statistics, Old Dominion University, Norfolk, Va.,
^cDepartment of Biostatistical Sciences, Section on Statistical Genetics, Wake Forest University Health Sciences,
Winston-Salem, N.C., ^dDepartment of Biostatistics, Section on Statistical Genetics and ^eClinical Nutrition
Research Center, University of Alabama at Birmingham, Birmingham, Ala., USA

Key Words

Multiple imputation · Measurement error · Admixture ·
Ancestry · Structured association testing

Abstract

Objectives: Structured association tests (SAT), like any statistical model, assumes that all variables are measured without error. Measurement error can bias parameter estimates and confound residual variance in linear models. It has been shown that admixture estimates can be contaminated with measurement error causing SAT models to suffer from the same afflictions. Multiple imputation (MI) is presented as a viable tool for correcting measurement error problems in SAT linear models with emphasis on correcting measurement error contaminated admixture estimates. **Methods:** Several MI methods are presented and compared, via simulation, in terms of controlling Type I error rates for both non-additive and additive genotype coding. **Results:** Results indicate that MI using the Rubin or Cole method can be used to correct for measurement error in admixture estimates in SAT linear models. **Conclusion:** Although MI can be used to correct for admixture measurement error in SAT linear models, the data should be of reasonable quality, in terms of marker informativeness, because the method uses the exist-

ing data to borrow information in which to make the measurement error corrections. If the data are of poor quality there is little information to borrow to make measurement error corrections.

Copyright © 2009 S. Karger AG, Basel

Introduction

In statistical modeling, ignoring confounding variables can lead to either increased false positive or increased false negative rates [1] and a bias in parameter estimates either away from or toward a null value. A confounder is a variable that is correlated with the predictor(s) and the outcome variable(s) in the model, and can cause a biased estimation of the causal association between these variables if not properly taken into account. To control for a confounder's effects, it is often included in the model as a covariate, which partials out its relationship with the predictor(s) and outcome variables in the model to obtain more accurate estimates of the relationship between predictor(s) and outcome(s) variables. In genetic association studies is overwhelming evidence that population stratification, assortative mating, and admixture among populations can result in intrapopulation varia-

tion in ancestry, correlations of allelic variation among unlinked loci, and ultimately confound association studies [2–6].

When discussing individual ancestry and individual admixture, it is important to distinguish what is meant by these two concepts. By individual ancestry (proportion) we mean the proportion of an individual's ancestors that come from a specified population. In contrast, individual admixture (proportion) is defined as the proportion of an individual's genome that is inherited from a specific parental population [7].

Several approaches to correct for population stratification and admixture have been proposed. Genomic control (GC) [4, 8, 9] and structured association testing (SAT) [10–13] are two such statistical approaches. Although GC can be useful in correcting for population stratification, we focus here on precisely estimating ancestry and using it as a covariate in SAT. The SAT model can flexibly accommodate time-to-event, dichotomous, ordinal, or continuous responses for the outcome measure and the model parameters can be estimated through standard statistical software. However, the model is subject to the same assumptions associated with standard linear models, including an implicit assumption that *all variables are measured without error*. In linear models, measurement error in predictors can introduce bias in the parameter estimates and increase the residual variance, which translates into inaccurate conclusions about hypotheses being tested.

Admixture may mask the true relationship between the phenotype (outcome variable) and genotypes (predictors) and produce false positives [14–17] and/or false negatives [18]. Individual admixture estimates are typically used as proxies for individual ancestry because individual ancestry is rarely known. Redden et al. [7] and Divers et al. [19] have shown that individual admixture estimates, as proxies for individual ancestry, are contaminated with measurement error for several reasons. First, only a subset of genetic markers with imperfectly known ancestral population allele frequencies is used to estimate admixture (i.e., not fully ancestry informative markers). Second, imperfect historical knowledge about the admixed population can lead to inaccurate estimates of individual admixture. Third, individual ancestry is the expected value of individual admixture, but the process of meiosis introduces random variation between the two constructs. Finally, genotyping errors will also contribute to individual admixture being estimated with error. All or any one of these conditions will cause a discrepancy between individual ancestry and estimates of individual

admixture, which translates into error contaminated ancestry estimates.

This paper addresses accounting for admixture measurement error in SAT and explores a specific alternative, multiple imputation (MI), to the methods previously described by Divers et al. [19]. We use simulation to evaluate the performance of the proposed methods and conclude with a discussion of results and how the methods can be extended.

Methods

SAT Model

Redden et al. [7] formulated SAT in the form of a general linear model as follows:

$$f(Y_i) = \beta_0 + \beta_1 A_i + \beta_2 P_{1i} P_{2i} + \beta_3 G_{ij1} + \beta_4 G_{ij2} + \varepsilon_i. \quad (1.1)$$

In the model $f(Y_i)$ is the link function linking Y_i variable (phenotype) to the parameters of the model, A_i is the ancestry of the i -th individual, P_{1i} and P_{2i} are the ancestry values of the two parents, and G_{ijk} is an indicator variable for the i -th individual with k and only k alleles at the j -th locus of type m (specific allele states). Redden et al. [7] propose inclusion of the product term for parental ancestry to better control for spurious association and achieve the desired Type I error rate. This general model can accommodate covariates such as gender, age, and treatment group and phenotypes such as time-to-event, dichotomous, ordinal, or continuous responses. The A_i ancestry component is included to control for the potential confounding effect and must either be assumed to be measured without error or necessitates a measurement error correction.

Ancestry and the Classical True Score Model

Admixture estimates can be expressed in the form of the classical true-score model (CTM) [20, 21] as

$$x_{ij} = \tau_i + u_{ij} \quad (1.2)$$

where x_{ij} is the j -th observed score (estimated admixture) for the i -th individual, τ_i is the true score (ancestry) for the i -th individual, and u_{ij} are the random components for the j -th admixture estimate ($j = 1, 2, \dots, p$). In the CTM it is typically assumed that $E(U_{ij}) = 0$ and $\text{var}(u_{ij}) = \sigma_u^2$ with u_{ij} mutually independent of each other and of τ_j [20, 21]. It can then be shown that $E(x_{ij}) = \tau_i$ or $\mu_{x_i} = \tau_i$ and $\sigma_x^2 = \sigma_\tau^2 + \sigma_u^2$. Note that τ_i and u_{ij} are latent variables that are never observed, but both influence x_{ij} , which is observed. Nevertheless, an estimate of σ_u^2 can be obtained using only the data from the x_{ij} 's. This can be done through a reliability coefficient, generically defined as

$$\sigma_{x\tau}^2 = \sigma_\tau^2 / \sigma_x^2 = \sigma_\tau^2 / (\sigma_\tau^2 + \sigma_u^2) \quad (1.3)$$

and ranges from 0 to 1 [20]. It should be noted that $\rho_{x\tau}^2$ is sometimes referred to as the intra-class correlation. Of specific interest here is Cronbach's alpha (α_c), a measure of the reliability of the sum of the equally weighted

$$x = \sum_{j=1}^p x_j$$

[22], computed as

$$\alpha_c = \left[p / (p - 1) \right] \left[1 - \frac{\sum_{j=1}^p \text{cov}(x_j, x_j)}{\sum_{j=1}^p \sum_{j'=1}^p \text{cov}(x_j, x_{j'})} \right]. \quad (1.4)$$

The computation of α_c only requires that the x_i 's measure the same construct or latent variable (i.e., Tau-equivalence) [22]. The estimated reliability coefficient in turn provides an estimate of σ_u^2 as $\sigma_u^2 = \sigma_x^2(1 - \rho_{x\tau}^2) = \sigma_x^2(1 - \alpha_c)$ [20] and is a weighted estimate of the observed score variance. Note that α_c is being used instead of $\rho_{x\tau}^2$. In genetic association/mapping studies of population data, ancestry informative markers (AIMs) on each of the autosomal chromosomes can be used to obtain chromosome-specific admixture estimate for each person who, conditional on true individual ancestry, is independent. From here on we denote admixture estimate for an individual by x_{ij} . The chromosome-specific admixture estimates can be used to estimate α_c . For a discussion of how Cronbach's alpha effects association tests see Divers et al. [19].

Linear Models with Measurement Error

Consider the linear model

$$Y = \beta_0 + \beta X + \varepsilon, \quad (1.5)$$

with $\varepsilon \sim NID(0, \sigma_\varepsilon^2)$. If X is measured with error, it can be shown that the Ordinary Least Squares (OLS) regression of Y on X yields a consistent estimator of

$$\beta^* = [\sigma_\tau^2 / (\sigma_\tau^2 + \sigma_u^2)] \beta = \alpha_c \beta, \quad (1.6)$$

which is attenuated towards zero. In addition, measurement error affects the residual variance as seen in the expression

$$\text{var}(Y | X) = \sigma_\varepsilon^2 + \sigma_u^2 [\sigma_\tau^2 / (\sigma_\tau^2 + \sigma_u^2)] \beta^2. \quad (1.7)$$

From the above two expressions, the smaller the measurement error variance (σ_u^2), the closer β^* will be to β and the residual variance will be less confounded. Of course, neither problem will exist when there is no measurement error ($\sigma_u^2 = 0$).

Divers et al. [19] demonstrated the use of quadratic measurement error correction (QMEC) [23, 24], regression calibration [25], expanded regression calibration [26, 27], and the simulation extrapolation (SimEx) algorithm [28, 29] to address the admixture measurement error challenge in SAT models. They found that the QMEC method performed best in terms of controlling the Type I error rate and the expanded regression calibration method performed the worst. However, the QMEC method is limited to linear models making a more flexible model desirable. Multiple imputation (MI) can in principle correct for measurement error in the general SAT model of Redden et al. [7] and flexibly accommodate a variety of special cases such as logistic and Cox regression.

Multiple Imputation for Measurement Error

Measurement error problems may be conceptualized as missing data problems in which we observe imperfect measurements but true scores are never seen (missing) [29]. Using MI to impute the missing true values as a means of correcting for measurement error in conjunction with alpha, which is used to estimate the measurement error variance, has the advantage of using the observed data as opposed to using (a) validation data in which the

true values of the variable are actually observed, (b) replication data where multiple measurements of the variable are made, or (c) instrumental data [29] in which two or more alternative methods are required to measure the variable.

In MI one treats imputed true values as probable and not as the one 'true' value, and using the one 'true' value ignores imputation variability or uncertainty about the actual value. Imputing a single value would fail to take into account the uncertainty about the actual value and can lead to underestimated standard errors, confidence intervals that cover less than their nominal coverage, and inflated Type I error rates. MI accounts for the uncertainty by imputing multiple values for each missing value and accounting for the resulting uncertainty and will yield valid estimates and tests pursuant to certain assumptions about the missing data mechanism [for details, see 32, 33].

Estimating True Scores

To use MI for measurement error correction one can proceed by obtaining an estimate of the true score (ancestry) for i -th individual based on the observed data [21] by formulating the prediction equation from regression theory as follows:

$$(\hat{Y}_i - \mu_Y) / \sigma_Y = \rho_{XY} (X_i - \mu_X) / \sigma_X, \quad (1.8)$$

where \hat{Y}_i is the predicted score, ρ_{XY} is the correlation between X and Y , μ_Y and μ_X , and σ_Y and σ_X are the means and standard deviations of Y and X , respectively. Equation 1.8 can be rewritten as

$$\hat{Y}_i = \rho_{XY} \frac{\sigma_Y}{\sigma_X} (X_i - \mu_X) + \mu_Y. \quad (1.9)$$

Substituting $\hat{\tau}_i$ for \hat{Y}_i , $\rho_{X\tau}$ for ρ_{XY} , $\sigma_\tau / \sigma_X = \rho_{X\tau}$, and $\mu_\tau = \mu_X$ yields

$$\hat{\tau}_i = \rho_{X\tau}^2 (X_i - \mu_X) + \mu_X. \quad (1.10)$$

Note that α_c is used instead of $\rho_{X\tau}^2$. The variance associated with this estimated true score is $\hat{\sigma}_u^2 = \hat{\sigma}_x^2 (1 - \hat{\alpha}_c)$. The reliability index is defined as $\rho_{X\tau} = \sigma_\tau / \sigma_X$ [21]. Equation 1.10 is a Bayesian or 'shrunken' estimator [30]. Thus, probable true scores can be generated using estimated coefficients ($\hat{\alpha}_c$) and variances ($\hat{\sigma}_u^2$). This idea will be revisited in the imputation process.

Implementing MI for Measurement Error Correction

Redden et al. [7] indicated that the product of parental ancestries is required to achieve the desired Type I error rate when genotypic (as opposed to simply allelic) effects at the marker locus are tested. Divers et al. [19] found that squaring the individual admixture estimate 'adequately approximates the product of ancestral ancestries'. Hence, in the present context, quadratic terms of the probable true scores are also required. Here, we justify the centering of the admixture estimate before implementation of MI. Assume that $X \sim N(\mu, \sigma^2)$, then

$$\text{cov}(X, X^2) = (\mu^3 + 3\mu\sigma^2) - \mu(\mu^2 + \sigma^2) \quad (1.11)$$

$$\text{cov}(X, X^2) = 2\mu\sigma^2. \quad (1.12)$$

By centering X , then $(X - \mu) \sim N(0, \sigma^2)$, it then follows that $\text{cov}((X - \mu), (X - \mu)^2) = 0$. Thus, centering the admixture estimate allows one to ignore the covariance between X and X^2 in the imputation process and subsequently only requires the squaring of the probable true score.

Using the SAT model proposed by Redden et al. [7], and given in equation (1.1), the following steps were implemented for the MI process.

1. Measurement model

a. *Regression method*: Regress X_i , the error contaminated variable, on the other variables in the model of interest. In our model this is $X_i = \beta_0 + \beta_Y Y_i + \beta_2 G_{ij,1} + \beta_3 G_{ij,2} + \varepsilon$. This step is identical to standard imputation routines in which X_i is the variable with missing values.

2. Imputation process: Draw regression coefficients from the posterior distribution

a. *Cole et al. (2006)*: This method uses the estimated parameters $\hat{\beta} = (\hat{\beta}_0 \hat{\beta}_Y \hat{\beta}_2 \hat{\beta}_3)'$ and $\hat{\Sigma}_{\hat{\beta}}$ from Step 1, where $\hat{\Sigma}_{(\cdot)} = \hat{\sigma}^2 (X'X)^{-1}$, and $\hat{\sigma}^2 = \alpha_c \hat{\sigma}_e^2$. Draw a new set of m random parameter estimates as $\tilde{\beta}^{(m)} = \hat{\beta} + V_{\hat{\beta}}' Z$ from Step 1, where $\hat{\Sigma}_{(\cdot)} = V_{(\cdot)} V_{(\cdot)}$, and Z is a vector of $z_i \sim NID(0, 1)$.

b. *Rubin (1987, pp 166–167)*: In this method draws are made from the new set of m random parameter estimates as $\tilde{\beta}^{(m)} = \hat{\beta} + \sigma_* V' Z$ from Step 1, where $(X'X)^{-1} = V'V$, $\sigma_*^2 = \hat{\sigma}^2 (df_e - 1)/g$, $\hat{\sigma}^2 = \alpha_c \hat{\sigma}_e^2$, $g \sim \chi^2 (df_e - 1)$, and df_e is the degrees of freedom (df) for the error term.

c. *Bootstrap (Rubin, 1987)*: With this method rather than making draws from $Z \sim NID(0, 1)$ as in 2(a) and 2(b), the residuals from the fitted model are bootstrapped. Everything remains the same as option 2(a) and 2(b) except that

$$e_i^* = e_i / \sqrt{\hat{\sigma}^2 (1 - k/n)}$$

is used instead of z_i , where e_i is the standardized residual for the i -th individual, $\hat{\sigma}^2$ is the estimated variance, k is the number of parameters in the model, and n is the sample size. This method has the advantage of imputing values whose distribution is similar to that of the observed values [31]. All options in Step 2 simulate draws from the posterior predictive distribution of the parameters. This allows for ‘proper’ imputation [32] because the estimates produced in Step 2 are only probable estimates and not the true estimates.

3. Imputation Process: Drawing m new probable true scores.

- a. $\tilde{T}_i^{(m)} = \tilde{\beta}_0^{(m)} + \tilde{\beta}_Y^{(m)} Y_i + \tilde{\beta}_2^{(m)} G_{ij,1} + \tilde{\beta}_3^{(m)} G_{ij,2} + z_i \hat{\sigma}$ (Cole)
- b. $\tilde{T}_i^{(m)} = \tilde{\beta}_0^{(m)} + \tilde{\beta}_Y^{(m)} Y_i + \tilde{\beta}_2^{(m)} G_{ij,1} + \tilde{\beta}_3^{(m)} G_{ij,2} + z_i \sigma_*$ (Rubin)
- c. $\tilde{T}_i^{(m)} = \tilde{\beta}_0^{(m)} + \tilde{\beta}_Y^{(m)} Y_i + \tilde{\beta}_2^{(m)} G_{ij,1} + \tilde{\beta}_3^{(m)} G_{ij,2} + e_i^* \sigma_*$ (Bootstrap),

where $z_i \sim NID(0, 1)$.

4. Fit the model of interest using the new m probable true scores. This is

$$Y_i^{(m)} = \beta_0 + \beta_1 \hat{T}_i^{(m)} + \beta_2 \hat{T}_i^{(m)2} + \beta_3 G_{ij,1} + \beta_4 G_{ij,2} + \varepsilon_i \quad (1.13)$$

for the SAT Model discussed.

5. Combine the m parameter estimates using the standard methods described by Rubin [31]. Additionally, adjusted df [33], which cannot exceed the complete-data df , were used to compute the df for MI inferences.

In the above steps, measurement correction is essentially variance correction in the form of $\hat{\sigma}^2 = \alpha_c \hat{\sigma}_e^2$.

It is important to recall that MI assumes that the missing values are missing at random (MAR). In short, MAR means the probability that values are missing on a certain variable Y depends on other variables in the model, but not on Y itself. Al-

though, MI is not specifically being used to impute missing values, the MAR assumption still holds. What is being treated as missing are the true value, which are not observed. Even so, it is assumed that the true values have a relationship with the other variables in the model, which is the MAR assumption.

For comparative purposes, the data were analyzed through a naïve model, a model that treats the variables as if they had no measurement error.

Simulation Study

The simulation investigated the effect of error-contaminated individual ancestry proportions on the Type I error rate in SAT models. The underlying individual ancestry distribution (X) was simulated by making draws from a mixture of uniform and normal distributions that mimic the ancestral distribution observed in African American populations following the simulation procedures by Tang et al. [34]. A thousand datasets, each containing 500 markers and 1000 individuals were generated. The delta-value of each marker is allowed to vary between 0 and 0.9. However, only ancestry informative markers were retained for individual ancestry proportion estimation. They were sampled more heavily toward the upper bound of this interval for high Cronbach’s alpha values and more toward the lower bound for lower Cronbach’s alpha values. These markers were evenly divided into 22 blocks, which are used to provide a set of 22 estimates of individual ancestry. These estimates are used to estimate Cronbach’s alpha. From these sets, 20 sets of 500 markers for each mean Cronbach Alpha values of $\bar{\alpha}_c = 0.90, 0.80, 0.70$ were randomly selected. The allele frequency of each marker in the admixed sample was computed as a mixture of two parental allele frequencies as follow:

$$P_{ij}^{adx} = X_i P_j^1 + (1 - X_i) P_j^2 \quad (1.14)$$

where P_j^1 and P_j^2 are frequencies of allele 1 at the j -th marker for the 1st and 2nd parental populations, X_i the simulated ancestry of the i -th admixed individual, and P_{ij}^{adx} is the allele 1 frequency for the i -th admixed individual for the j -th marker. In this simulation, given a specific delta value, $P_j^1 \sim U(0, 1)$, $P_j^2 = P_j^1 + \delta$ where $\delta \sim Bin(100, delta) \times 0.01$, and $X_i = 0.2 \times U(0.1, 0.9) + 0.8 \times N(0.15, 0.05^2)$ [19, 34]. The trait or phenotypic variable was generated as

$$Y_i = 35 + 5X_i + 0G_{ij,1} + 3G_{ij,2} + \varepsilon_i \quad (1.15)$$

$$Y_i = 35 + 5X_i + 5X_i^2 + 0G_{ij,1} + 3G_{ij,2} + \varepsilon_i \quad (1.16)$$

for the linear and quadratic model, respectively, where $\varepsilon_i \sim N(0, 4)$. The linear model was generated for comparative purposes. In the simulation X_i is the simulated true ancestry proportion from the above mixture distribution and $W_i = X_i + e_i$ is the observed ancestry proportion, where $e_i \sim N(0, \sigma_i^2)$, is the error-contaminated ancestry coefficient. Note that this is ancestry estimated in the form of the classical true-score model (CTM). The σ_i^2 values were selected so that the observed correlations between W_i and X_i vary between 0.85 and 0.95, and to demonstrate that highly yet still imperfectly correlated true and estimated (or measured) ancestry proportions can still lead to Type I error inflation. We note that a correlation between 0.85 and 0.95 ensures that Cronbach’s alpha is bounded between 0.7 and 0.9. Under this scheme, 20 datasets of 500 markers containing 1000 individuals were simulated for a total of 10,000 markers. Each marker was tested for association with the simulated phenotype.

Analysis of Simulation

Each dataset contained a sample of 1000 individuals with 500 markers. Both the SAT models with and without the squared ancestry term were fitted to the data; we refer to the former as a linear SAT model and the latter a quadratic SAT model. Assume there are two alleles (A, a) at a locus forming three genotypes (aa, aA, AA) and allele A is of interest. The genotypes can be coded to allow for testing of only additive or both additive and non-additive effects and table 1 offers respective coding schemes.

Results

Table 2 contains the Type I error rates of the linear and quadratic SAT models with additive and non-additive genotypic coding for different reliability coefficients' corresponding to naïve model (i.e. without measurement correction). The type I error rates are liberal irrespective of genotype coding, a linear or quadratic SAT model, and reliability coefficient, implying that the association test will have a higher false positive rate if there is confounding by admixture and the model is not corrected for measurement error.

Tables 3–5 provide the type I error rates with measurement correction corresponding to the Rubin, Bootstrap,

Table 1. Coding of genotypic values in simulation of genetic data

Genotype	Non-additive		Additive model
	$G_{(aA)i}$	$G_{(AA)i}$	$G_{(A)i}$
aa	0	0	0
aA	1	0	1
AA	0	1	2

Table 2. Type I error rates corresponding to the β coefficients in SAT models without any measurement corrections

SAT model	Genotype coding	Parameter	Reliability coefficient		
			0.90	0.80	0.70
Linear	additive	β_3	0.0758	0.0848	0.1830
	non-additive	β_3	0.0702	0.0824	0.1684
		β_4	0.0596	0.0632	0.0910
Quadratic	additive	β_3	0.0876	0.1002	0.2208
	non-additive	β_3	0.0736	0.0912	0.2054
		β_4	0.0574	0.0688	0.1082

Table 3. Average Type I error rates after measurement correction corresponding to the β coefficients for reliability coefficient of 0.90, using 10,000 replicates

SAT model	Genotype coding	Method	Parameter	Number of imputations		
				5	10	15
Linear	additive	Rubin	β_3	0.0294	0.0288	0.0266
		Bootstrap	β_3	0.0400	0.0430	0.0446
		Cole	β_3	0.0288	0.0290	0.0268
	non-additive	Rubin	β_3	0.0276	0.0262	0.0252
			β_4	0.0308	0.0300	0.0306
		Bootstrap	β_3	0.0384	0.0376	0.0348
			β_4	0.0364	0.0368	0.0386
		Cole	β_3	0.0306	0.0268	0.0302
			β_4	0.0328	0.0320	0.0322
Quadratic	additive	Rubin	β_3	0.0386	0.0364	0.0336
		Bootstrap	β_3	0.0500	0.0470	0.0454
		Cole	β_3	0.0406	0.0384	0.0386
	non-additive	Rubin	β_3	0.0382	0.0374	0.0404
			β_4	0.0290	0.0302	0.0258
		Bootstrap	β_3	0.0672	0.0682	0.0662
			β_4	0.0526	0.0522	0.0528
		Cole	β_3	0.0430	0.0406	0.0434
			β_4	0.0312	0.0292	0.0328

Table 4. Average Type I error rates after measurement correction corresponding to the β coefficients for reliability coefficient of 0.80, using 10,000 replicates

SAT model	Genotype coding	Method	Parameter	Number of imputations		
				5	10	15
Linear	additive	Rubin	β_3	0.0332	0.0322	0.0356
		Bootstrap	β_3	0.0482	0.0484	0.0498
		Cole	β_3	0.0388	0.0314	0.0356
	non-additive	Rubin	β_3	0.0314	0.0300	0.0308
			β_4	0.0336	0.0324	0.0320
		Bootstrap	β_3	0.0478	0.0478	0.0466
			β_4	0.0424	0.0400	0.0382
		Cole	β_3	0.0316	0.0322	0.0322
			β_4	0.0356	0.0316	0.0346
Quadratic	additive	Rubin	β_3	0.0460	0.0478	0.0454
		Bootstrap	β_3	0.0698	0.0678	0.0660
		Cole	β_3	0.0504	0.0452	0.0480
	non-additive	Rubin	β_3	0.0440	0.0450	0.0442
			β_4	0.0352	0.0332	0.0332
		Bootstrap	β_3	0.0944	0.0964	0.0988
			β_4	0.0622	0.0612	0.0588
		Cole	β_3	0.0464	0.0494	0.0452
			β_4	0.0358	0.0328	0.0330

and Cole methods. Table 3 contains Type I error rates for both the linear and quadratic SAT model with additive and non-additive genotypic coding for reliability coefficient of 0.90. The type I error rates for all three methods of imputation were slightly conservative for the linear SAT model. A similar trend occurred for the quadratic SAT model with the exception of the bootstrap method, where the type I error rates for the β_3 were slightly liberal.

The Type I error rates of the linear and quadratic SAT model with additive and non-additive genotypic coding with reliability coefficient of 0.80 are presented in table 4 with measurement correction using the Rubin, Bootstrap, and Cole's method. For the linear SAT model, the Bootstrap imputation method controlled the Type I error rate best followed closely by the Cole and Rubin's methods. Additionally, the Cole and Rubin methods were not as conservative as before. However, the type I error rates were liberal for the quadratic SAT model using the Bootstrap method irrespective of genotype coding system. Both Rubin and Cole's methods provided type I error rates closer to nominal significance level of 0.05 and slightly less conservative compared to the situation with reliability coefficient of 0.90.

Lastly, table 5 displays the Type I error rates of the linear and quadratic SAT models with additive and non-additive genotypic coding with reliability of 0.70. The type I error rates for the Bootstrap method were very liberal compared to either of Rubin's or Cole's method. However, all methods performed poorly for the quadratic SAT model. However, the Rubin and Cole methods kept the type I error rate closer the nominal significance level of 0.05. The slight exception here is that both the Rubin and Cole methods were slightly conservative for the β_4 parameter estimate.

Discussion

Measurement error in linear model variables is an important consideration, and through simulation we demonstrated the importance for correcting measurement error in linear models. Of particular interest was using multiple imputation (MI) for measurement error correction for the Redden et al. [7] SAT model. Although the Redden SAT model requires individual ancestry estimates to control for admixture confounding, individual admixture estimates were used because individual ancestry estimates are rarely known, so admixture estimates

Table 5. Average Type I error rates after measurement correction corresponding to the β coefficients for reliability coefficient of 0.70, using 10,000 replicates

SAT model	Genotype coding	Method	Parameter	Number of imputations		
				5	10	15
Linear	additive	Rubin	β_3	0.0480	0.0510	0.0522
		Bootstrap	β_3	0.1214	0.1262	0.1248
		Cole	β_3	0.0500	0.0516	0.0536
	non-additive	Rubin	β_3	0.0444	0.0412	0.0444
			β_4	0.0400	0.0356	0.0366
		Bootstrap	β_3	0.1138	0.1138	0.1188
			β_4	0.0636	0.0648	0.0666
		Cole	β_3	0.0436	0.0464	0.0420
			β_4	0.0374	0.0382	0.0384
Quadratic	additive	Rubin	β_3	0.0926	0.0972	0.0990
		Bootstrap	β_3	0.2218	0.2324	0.2350
		Cole	β_3	0.0928	0.0994	0.0984
	non-additive	Rubin	β_3	0.0908	0.0904	0.0932
			β_4	0.0576	0.0524	0.0558
		Bootstrap	β_3	0.2328	0.2462	0.2496
			β_4	0.1106	0.1184	0.1130
		Cole	β_3	0.0934	0.0998	0.1018
			β_4	0.0586	0.0582	0.0560

can be used as a surrogate for the ancestry estimates. We then describe how to use MI for measurement correction. Like Divers et al. [19], we also used Cronbach's alpha [35] as a component of our measurement error correction procedure. We also described three different methods for imputing probable true scores for admixture: Rubin, Bootstrap, Cole.

In the linear SAT model, of the three different methods for imputing probable admixture scores, the Rubin and Cole methods appear to work best. Although at first it looks like the Bootstrap method controls the Type I error correctly whereas the Rubin and Cole methods are slightly conservative, as the marker informativeness begins to decrease it is the Rubin and Cole methods that control Type I error rate and the Bootstrap method becomes liberal. Consistently, the Rubin and Cole method provided better control of the Type I error rate than the Bootstrap method. This same pattern was observed in Divers et al. [19], in that measurement error correction only appears to be required when the informativeness of the markers is of intermediate value. The reason for this is that when markers are highly informative, the measurement correction method provides little improvement. On the other hand, when marker informativeness is low, the measurement correction method has poor in-

formation to borrow for measurement correction. MI for measurement correction as presented uses the existing data to accomplish this goal and require no external information.

In the quadratic SAT model, of the three different methods for imputing probable admixture scores, the Rubin and Cole methods again appear to work best. The Bootstrap method did not consistently provide reasonable control of the Type I error rate. One interesting point is that the type I error rates of the Bootstrap method, in all models, are very similar to the type I error rates of the model without measurement error correction, suggesting that the Bootstrap method is not providing much measurement error correction. Notably, none of the methods works particularly well for a quadratic SAT model with admixture reliability of 0.70. Because of this result the linear SAT model corrected for measurement error may be considered, yet it too can have problems if the genetic effects are markedly non-additive (e.g., overdominance).

There is now much agreement that population admixture and/or population stratification can confound association studies when not taken into account. However, it should also be mentioned that accuracy with which admixture is measured will have an influence on Type I error. When admixture or any other continuous variable

are contaminated with error, MI for measurement error correction can help control the specified Type I error rate. However, this method is only useful if the data are of reasonably good quality with respect to marker information, which means that much care should still be taken when designing association studies, and in particular when measuring variables that will be used for analysis in a statistical model.

Acknowledgements

This work was supported in part by National Institutes of Health grants: 5R01AR052658-02, ES009912, DK056336, CA100949-03, HL072757, AR007450, AR049084, R21LM008791, R01GM077490. The opinions expressed are solely those of the authors and do not necessarily represent those of the NIH or any other organization with which the authors are affiliated.

References

- Weinberg CR: Toward a clearer definition of confounding. *Am J Epidemiol* 1993;137: 1–8.
- Knowler WC, Williams RC, Pettitt DJ, Steinberg AG: Gm3;5,13,14 and type 2 diabetes mellitus: An association in american indians with genetic admixture. *Am J Hum Genet* 1988;43:520–526.
- 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.
- Devlin B, Roeder K: Genomic control for association studies. *Biometrics* 1999;55:997–1004.
- Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, Gabriel SB, Topol EJ, Smoller JW, Pato CN, Pato MT, Petryshen TL, Kolonel LN, Lander ES, Sklar P, Henderson B, Hirschhorn JN, Altshuler D: Assessing the impact of population stratification on genetic association studies. *Nat Genet* 2004;36:388–393.
- Redden DT, Allison DB: The effect of assortative mating upon genetic association studies: Spurious associations and population substructure in the absence of admixture. *Behav Genet* 2006;36:678–686.
- Redden DT, Divers J, Vaughan LK, Tiwari HK, Beasley TM, Fernández JR, Kimberly RP, Feng R, Padilla MA, Liu N, Miller MB, Allison DB: Regional admixture mapping and structured association testing: Conceptual unification and an extensible general linear model. *Plos Genetics* 2006;2:1254–1264.
- Devlin B, Bacanu SA, Roeder K: Genomic control to the extreme. *Nat Genet* 2004;36: 1129–1130;author reply 1131.
- Devlin B, Roeder K, Wasserman L: Genomic control, a new approach to genetic-based association studies. *Theor Popul Biol* 2001;60: 155–166.
- Pritchard JK, Donnelly P: Case-control studies of association in structured or admixed populations. *Theor Popul Biol* 2001;60:227–237.
- Satten GA, Flanders WD, Yang Q: Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *Am J Hum Genet* 2001;68:466–477.
- Chen HS, Zhu X, Zhao H, Zhang S: Qualitative semi-parametric test for genetic associations in case-control designs under structured populations. *Ann Hum Genet* 2003;67: 250–264.
- Zhang S, Zhu X, Zhao H: On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals. *Genet Epidemiol* 2003; 24:44–56.
- Ziv E, Burchard EG: Human population structure and genetic association studies. *Pharmacogenomics* 2003;4:431–441.
- Hoggart CJ, Parra EJ, Shriver MD, Bonilla C, Kittles RA, Clayton DG, McKeigue PM: Control of confounding of genetic associations in stratified populations. *Am J Hum Genet* 2003;72:1492–1504.
- Halder I, Shriver MD: Measuring and using admixture to study the genetics of complex diseases. *Hum Genomics* 2003;1:52–62.
- Marchini J, Cardon LR, Phillips MS, Donnelly P: The effects of human population structure on large genetic association studies. *Nat Genet* 2004;36:512–517.
- Deng HW: Population admixture may appear to mask, change or reverse genetic effects of genes underlying complex traits. *Genetics* 2001;159:1319–1323.
- Divers J, Vaughan LK, Padilla MA, Fernandez JR, Allison DB, Redden DT: Correcting for measurement error in individual ancestry estimates in structured association tests. *Genetics* 2007;176:1823–1833.
- Allen MJ, Yen WM: Introduction to measurement theory. Monterey, CA, Brooks/Cole Pub. Co., 1979.
- Crocker LM, Algina J: Introduction to classical and modern test theory. New York, Holt, Rinehart, and Winston, 1986.
- Bollen KA: Structural equations with latent variables. New York, Wiley, 1989.
- Cheng C-L, Van Ness JW: Statistical regression with measurement error. London, Arnold, 1999.
- Cheng CL, Schneeweiss H: Polynomial regression with errors in the variables. *J R Stat Soc Ser B (Statistical Methodology)* 1998;60: 189–199.
- Carroll RJ, Stefanski LA: Approximate quasi-likelihood estimation in models with surrogate predictors. *J Am Stat Ass* 1990;85: 652–663.
- Schneeweiss H, Nitter T: Estimating a polynomial regression with measurement errors in the structural and in the functional case – a comparison; in Mohammed AK, Saleh E (eds): *Data Analysis from Statistical Foundations: A Festschrift in Honour of the 75th Birthday of Das Fraser*. Huntington, NY, Nova Science Publishers, 2001, pp 195–207.
- Kuha J, Temple J: Covariate measurement error in quadratic regression. *Int Stat Rev* 2003; 71:131–150.
- Cook JR, Stefanski LA: Simulation-extrapolation estimation in parametric measurement error models. *J Am Stat Ass* 1994;89: 1314–1328.
- Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM: *Measurement error in nonlinear models: A modern perspective*, ed 2. Boca Raton, Chapman & Hall/CRC, 2006.
- Lindley DV, Smith AFM: Bayes estimates for the linear model. *J R Stat Soc* 1972;34:1–41.
- Rubin DB: *Multiple imputation for nonresponse in surveys*. New York, Wiley, 1987.
- Little RJA, Rubin DB: *Statistical Analysis with Missing Data*, ed 2. Hoboken, NJ, Wiley-Interscience, 2002.
- Barnard J, Rubin DB: Small-sample degrees of freedom with multiple imputation. *Biometrika* 1999;86:948–955.
- Tang H, Peng J, Wang P, Risch NJ: Estimation of individual admixture: Analytical and study design considerations. *Genet Epidemiol* 2005;28:289–301.
- Cronbach LJ: Coefficient alpha and the internal structure of tests. *Psychometrika* 1951; 16:297–334.