UW Biostatistics Working Paper Series

January 2006

Comparison of Haplotype-based and Tree-based SNP Imputation in Association Studies

James Y. Dai
University of Washington, yud@u.washington.edu

Ingo Ruczinski
Johns Hopkins University, ingo@jhu.edu

Michael LeBlanc
Fred Hutchinson Cancer Research center, mikel@crab.org

Charles Kooperberg
Fred Hutchinson Cancer Research Center, clk@fhcrc.org

Suggested Citation

Dai, James Y.; Ruczinski, Ingo; LeBlanc, Michael; and Kooperberg, Charles, "Comparison of Haplotype-based and Tree-based SNP Imputation in Association Studies" (January 2006). *UW Biostatistics Working Paper Series.* Working Paper 278.
https://biostats.bepress.com/uwbiostat/paper278

Introduction

It is widely recognized that complex diseases are likely caused by multiple susceptible loci, each contributing a small to medium amount to the disease risk, that are potentially interacting with each other [Risch, 1999; Risch, 2000; Botstein and Risch, 2003]. While linkage analysis shows to be largely ineffective, association studies, in which the frequencies of marker alleles in affected individuals and controls (either population- or family-based) are compared, may hold the promise of dissecting the genetic susceptibility of complex diseases [Risch, 2000; Botstein and Risch, 2003]. With the explosion of single nucleotide polymorphism (SNP) discovery and the advances in genotyping technologies, numerous SNP-based association studies have been carried out in a scale ranging from a few candidate genes to the whole genome [Barnby et al., 2005; Cope et al., 2005; Hu et al., 2005]. Despite the much improved cost efficiency in genotyping, missing data are fairly common in these association studies, often with a rate of 5% – 10%.

Depending on the analytical strategy undertaken, the missing SNPs have different impact on association inference. The haplotype approach treats a collection of adjacent SNPs in linkage disequilibrium (LD) all together and models the disease-haplotype association [Schaid et al., 2002; Zhao et al., 2003; Epstein et al., 2003; Stram et al., 2003]. Missing SNPs are essentially imputed in the process of haplotype reconstruction. The haplotype ambiguity introduced by missing SNPs costs enlarged variances in the estimated haplotype effects. Although haplotype analysis is effective to model interactions between SNPs in a tight LD block, it runs into difficulties when there are a large number of SNPs under investigation (a genome-wide scanning study, for example) and LD blocks are not well defined. In view of the polygenic nature of complex diseases, an alternative strategy is to directly regress disease status on the SNP main effects and SNP-SNP interactions. Cordell and Clayton [2002] proposed a stepwise logistic regression procedure for both case control data and family data. Ruczinski et al. [2003] developed logic regression, an adaptive regression methodology well suited for detecting interactions between binary SNP variables. These SNP-based approaches build the regression model by search algorithms, offer a flexible choice of hypothesis testing, yet remain computationally tractable. However, missing data in SNP genotypes pose a more serious problem to regression approaches.

The standard procedure to cope with the missing SNPs is to ignore the individuals that have missing values in the SNP loci under investigation, the so-called complete-case analysis. In general, the complete-case analysis reduces the effective sample size and potentially introduces bias in parameter estimates [Greenland and Finkle, 1995]. In particular, if a large number of SNPs are under investigation simultaneously (as in regression approaches), the proportion of individuals with at least one missing value can be quite high, even if the rate of missing SNPs is low for each locus. Given that neighboring SNPs are likely in linkage disequilibrium, it is feasible to impute missing ones by borrowing information from the observed ones. Furthermore, the imputation may also benefit from incorporating information on disease status and covariates. For example, when we studied the association between breast cancer and polymorphisms in the XPD gene in a matched case control study, the imputation frequencies for missing SNPs relied strongly on disease status and whether there was a family history of breast cancer [Brewster et al., 2005]. It is therefore desirable to develop a flexible imputation approach which takes into account LD in neighboring SNPs, as well as disease status and covariates if they are relevant.

The aforementioned haplotype reconstruction can be used for imputation. Existing EM algorithms [Excoffier and Slatkin, 1995; Qin et al., 2002] accommodate missing SNPs by first replacing the missing locus with all possible alleles. After haplotype reconstruction, the missing SNP genotypes are filled by sampling compatible haplotypes from their conditional distributions given the unphased genotypes. Similarly Bayesian methods for haplotype reconstruction can be use to impute the missing SNPs [Stephens et al., 2001; Niu et al., 2002; Lin et al, 2002]. All these methods may over-simplify the haplotype distribution in case control samples, as the frequencies of the disease-associated haplotypes may differ between cases and controls. To alleviate this problem, Lake et al. [2003] used a weighted EM (WEM) algorithm to jointly model the haplotype effects and haplotype frequencies. Alternatively, Epstein and Satten [2003] developed a retrospective likelihood, and estimated haplotype frequencies separately in cases and in controls. These more sophisticated methods can be easily adapted to imputing missing SNPs, with disease status and extra covariates being accounted for.

Instead of using genetic models, nonparametric regression methods such as classification and regression trees (CART) [Breiman et al., 1984] can be used to model the missing SNPs. Recently there is growing interest in applying tree methods to genetic association studies with a large number of SNPs [Zhang and Bonney, 2000; Bureau et al., 2005]. Previous applications of CART mostly target the association between SNPs and diseases, where the disease status is treated as an outcome variable. For the imputation purpose, we can regress each SNP locus with missing data on the other SNP loci, the covariates and the disease status, build the tree and predict the missing data at the locus. In order to obtain the joint distribution of missing SNPs at different loci, we employ a Gibbs sampler which iteratively cycles the regression and prediction by CART through loci with missing SNPs. One advantage of CART is that it deals with missing data by surrogate splits. That is, after choosing the best predictor and split point using the available data, a list of surrogate variables and split points are formed by comparing the performance of the alternate predictor with the primary predictor. If a primary predictor is missing for one individual, we use the secondary predictor if available, and so on.

In this article we develop and compare the haplotype and CART based imputation approaches in SNP association studies. In particular, we consider the EM and WEM algorithm as two representatives of haplotype-based approaches because of their easy implementation. We choose the case-control design as an illustrative example since it is probably the most commonly used in association studies. By comparing the imputation accuracy, bias and efficiency in inference, we evaluate the potential of the tree based approach as compared to the haplotype based approaches, and assess the benefit of the weighted EM approach as opposed to the regular EM approach. Most of all, we aim to demonstrate the benefit of a reasonable imputation strategy over simply ignoring the missing data.

Methods

Assume we have a case control study with $i = 1, 2, \ldots, n$ unrelated individuals. Let $\mathbf{D}_i = 1$ if individual i is a case and $\mathbf{D}_i = 0$ otherwise, and let $\mathbf{G}_i = (g_{i1}, g_{i2}, \ldots, g_{iK})$ be the unphased SNP data on individual i at K loci of interest. Some of the g_{ik} may be missing. Assume that in the population there are m possible haplotypes $h_1, h_2, \ldots, h_m$ with (unknown) population frequencies $\mathbf{p} = (p_1, p_2, \ldots, p_m)$. In addition to the genetic information we also have information on r covariates $\mathbf{X}_i = (x_{i1}, x_{i2}, \ldots, x_{ir})$.

Haplotype-based imputation

Treating haplotypes $\mathbf{H}_i = (h_{l(i)}, h_{l'(i)})$ as missing data, the EM algorithm [Excoffier and Slatkin, 1995] aims to maximize the likelihood

$$\prod_{i=1}^{n} \Pr(\mathbf{G}_i | p_1, p_2, \ldots, p_m) = \prod_{i=1}^{n} \sum_{\mathbf{H}_i \in \mathcal{G}_i} p_{l(i)} p_{l'(i)},$$

where $l(i)$ refers to a conformable haplotype to the observed $\mathbf{G}_i$, and $\mathcal{G}_i$ is the set of all possible haplotype pairs that conform to the observed $\mathbf{G}_i$. If the SNP at k locus is missing for individual i, all possible genotypes at k locus are filled in to construct conformable haplotypes. In the E step, the conditional probability of each pair of conformable haplotypes to unphased genotypes is calculated based on the current estimates of the haplotype frequencies,

$$\Pr((h_1, h_2) | \mathbf{G}_i) = \frac{\widehat{p}_1 \widehat{p}_2}{\sum_{\mathbf{H}_i \in \mathcal{G}_i} \widehat{p}_{l(i)} \widehat{p}_{l'(i)}}. \tag{1}$$

Note that Hardy-Weinberg equilibrium assumes that $\Pr(h_{l(i)}, h_{l'(i)}) = p_{l(i)} p_{l'(i)}$. The frequency estimates are then re-estimated in the M-step. At convergence,we use the conditional probabilities of all conformable haplotype pairs in (1) to impute the missing SNPs.

The weighted EM (WEM) approach is an extension of the EM algorithm which incorporates disease status, as haplotype frequencies may be different between cases and controls [Lake et al., 2003], as well as other covariates that may affect the disease risk and haplotype frequencies. Given $\mathbf{H}_i$ and $\mathbf{X}_i$ we model the disease penetrance by a logistic function

$$\Pr(\mathbf{D}_i = 1|\mathbf{H}_i = (h_{l(i)}, h_{l'(i)}), \mathbf{X}_i) = \frac{\exp[\alpha + \mathbf{1}(h_{l(i)}, h_{l'(i)})\boldsymbol{\gamma} + \mathbf{X}_i\boldsymbol{\beta}]}{1 + \exp[\alpha + \mathbf{1}(h_{l(i)}, h_{l'(i)})\boldsymbol{\gamma} + \mathbf{X}_i\boldsymbol{\beta}]}, \qquad (2)$$

where $\mathbf{1}(h_{l(i)}, h_{l'(i)})$ denotes a length m indicator vector. For simplicity, we assumes an additive model so that the two elements of $\mathbf{1}(h_{l(i)}, h_{l'(i)})$ equal to 1 and all other elements are 0. For a homozygous individual, the $l(i)^{th}$ element of $\mathbf{1}(h_{l(i)}, h_{l'(i)})$ equals 2. Our interest is not to use (2) to model the haplotype-disease association, but rather we use it as a vehicle to impute the missing SNPs.

Set $\boldsymbol{\Theta} = (\boldsymbol{\gamma}, \boldsymbol{\beta}, \mathbf{p})$. The complete data log-likelihood can be written as

$$\ell(\boldsymbol{\Theta}; \mathbf{D}, \mathbf{G}|\mathbf{X}) = \sum_{i=1}^{n} \left[\ell(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{D}_i|\mathbf{G}_i, \mathbf{X}_i) + \ell(\mathbf{p}; \mathbf{G}_i|\mathbf{X}_i)\right].$$

Since $\mathbf{H}_i$ has a finite number of possible values for a given $\mathbf{G}_i$, we sum this log-likelihood over all possible haplotypes in the Expectation step. Let

$$\begin{aligned} Q_i(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(s)}) &= \sum_{i=1}^{n} \mathrm{E}_{\mathbf{H}_i|\mathbf{D}_i,\mathbf{G}_i,\mathbf{X}_i,\boldsymbol{\Theta}^{(s)}}[\ell(\boldsymbol{\Theta}; \mathbf{D}_i, \mathbf{G}_i|\mathbf{X}_i)] \\ &= \sum_{i=1}^{n} \sum_{\mathbf{H}_i \in \mathcal{G}_i} W_{i,(s)} \left[\ell(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{D}_i|\mathbf{H}_i, \mathbf{X}_i) + \ell(\mathbf{p}; \mathbf{H}_i|\mathbf{X}_i)\right], \end{aligned}$$

where $\boldsymbol{\Theta}^{(s)}$ denotes the parameter estimates in the s^{th} iteration of the algorithm, $\boldsymbol{\Theta}$ denotes the parameter in the $(s+1)^{th}$ iteration and $W_{i,(s)} = \Pr(\mathbf{H}_i|\mathbf{D}_i, \mathbf{G}_i, \mathbf{X}_i, \boldsymbol{\Theta}^{(s)})$ is the conditional probability of a haplotype pair given the observed data and the current estimates of parameters. Note the first part of expected log likelihood is a weighted log likelihood for a generalized linear model, such as logistic regression in (2) if case control data. The second part is a weighted multinomial log likelihood. Both can be readily maximized using existing software. It is straightforward to calculate $W_{i,(s)}$ using Bayes Theorem

$$W_{i,(s)} = \frac{\Pr(\mathbf{D}_i|\mathbf{H}_i, \mathbf{X}_i, \boldsymbol{\Theta}^{(s)})\Pr(h_{l(i)}|\mathbf{X}_i, \boldsymbol{\Theta}^{(s)})\Pr(h_{l'(i)}|\mathbf{X}_i, \boldsymbol{\Theta}^{(s)})}{\sum_{\mathbf{H}_i \in \mathcal{G}_i} \Pr(\mathbf{D}_i|\mathbf{H}_i, \mathbf{X}_i, \boldsymbol{\Theta}^{(s)})\Pr(h_{l(i)}|\mathbf{X}_i, \boldsymbol{\Theta}^{(s)})\Pr(h_{l'(i)}|\mathbf{X}_i, \boldsymbol{\Theta}^{(s)})}. \qquad (3)$$

The derivation assumes Hardy-Weinberg equilibrium. Here $\Pr(\mathbf{D}_i|\mathbf{H}_i, \mathbf{X}_i, \boldsymbol{\Theta}^{(s)})$ is the current estimate of (2). The frequencies of haplotypes may depend on the covariates. If not,

$\Pr(h_{l(i)}|\mathbf{X}_i, \mathbf{\Theta}^{(s)})$ reduces to $p^{(s)}_{l(i)}$. We impute the missing SNPs by sampling the conformable haplotype pairs according to (3) at convergence. Although we describe the WEM approach for a case-control study with logistic regression, in principle it works for other generalized linear models.

The implementation of the EM and WEM algorithms is an adaptation of the existing **R** package *HaploStats* [Schaid et al., 2002; Lake et al., 2003]. We applied the *haplo.em* function to perform the EM algorithm. Rather than the regular EM algorithm, this function uses an efficient algorithm which progressively inserts a batch of SNP loci, enumerates possible haplotypes, runs EM, and trims off haplotypes with conditional probabilities below a threshold. We set the batch size to be 3, and the minimal conditional probability to 0.001. Starting from the *haplo.em* function, we develop a weighted EM algorithm similar to the *haplo.glm* function in *HaploStats*. The minimum haplotype frequency allowed is set to 10^{-6}.

Tree-based imputation

The tree-based approach is a general algorithm to impute the missing data, including missing SNPs and missing covariates in SNP association studies. For each individual i, let $\mathbf{M}_i = (M_{i1}, M_{i2}, \ldots, M_{ip})$ be the vector of p variables consisting of the covariates $\mathbf{X}_i = (x_{i1}, \ldots, x_{ir})$ and the unphased SNP data $\mathbf{G}_i = (g_{i1}, \ldots, g_{iK})$ which have missing entries $(1 \leq p \leq m+K)$. Let $\mathbf{C}_i$ be the vector of the remaining covariates for which all data are available. We assume that the outcome $\mathbf{D}_i$ is always observed. The joint probability distribution of the missing data for individual i given the observed data, $\Pr(M_{i1}, M_{i2}, \ldots, M_{ip}|\mathbf{C}_i, \mathbf{D}_i)$, is difficult to get. An obvious problem is that the sets of missing data $\mathbf{M}_i$ and complete data $\mathbf{C}_i$, respectively, are different for each individual i. Instead of modeling the joint distribution, we use the Gibbs sampler, a Markov chain Monte Carlo technique that uses conditional (low-dimensional) distributions to draw samples from a high-dimensional distribution.

Specifically, we consider iteratively sampling from the following sequence of the full conditional distributions in the $(n+1)^{th}$ iteration:

$$\begin{aligned}
M_1^{(n+1)} &\sim \Pr(M_1|M_2^{(n)}, M_3^{(n)}, \ldots, M_p^{(n)}, \mathbf{C}, \mathbf{D}) \\
M_2^{(n+1)} &\sim \Pr(M_2|M_1^{(n+1)}, M_3^{(n)}, \ldots, M_p^{(n)}, \mathbf{C}, \mathbf{D}) \\
&\vdots \\
M_p^{(n+1)} &\sim \Pr(M_p|M_1^{(n+1)}, M_2^{(n+1)}, \ldots, M_{p-1}^{(n+1)}, \mathbf{C}, \mathbf{D}).
\end{aligned}$$

where each full conditional distribution, for example $\Pr(M_1|M_2^{(n)}, M_3^{(n)}, \ldots, M_p^{(n)}, \mathbf{C}, \mathbf{D})$, is modeled by CART. This is easily done even though $M_2, \ldots, M_p$ contain missing observations before imputation has taken place, as CART uses surrogate splits if missing observations are encountered in a node [Breiman et al., 1984]. For example, if M_1 are actual data from a SNP, each terminal leaf in the classification trees provide a multinomial distribution from which we can sample. A convenient property of surrogate splits is that we do not have to guess the initial values of the missing data in $\mathbf{M}$, as a result only a very short burn-in of the above sampler is required. Under mild regularity conditions, this sequence of conditional variables converge to the joint distribution of missing data.

A similar idea, data augmentation [Tanner and Wong, 1997], has been exploited to deal with missing data in a Bayesian framework. However data augmentation is only analytically tractable in some simple situations, such as a multivariate normal distribution. The advantage of applying decision trees such as CART [Breiman et al., 1984] is that it can handle variables of any type, such as the 3-level factor (0,1,2) coding SNP genotypes in a locus, or a continuous age variable. Though lacking a formal proof, it has been demonstrated in simulation studies that the inference in missing data problems is fairly non-sensitive to model misspecification as long as the distribution of the missing data given the observed data involves the covariates that are ultimately found to be important in the model [Schafer, 1997]. It is therefore natural to investigate the performance of nonparametric regression methods such as decision trees for imputation. This has been suggested in the literature before [Harrell, 2001], though not for SNP association studies.

Our tree algorithm is based on the *rpart* package in **R** [Therneau and Atkinson, 1997]. The nodes in the decision trees generated by this package are split until the improvement of impurity measure (by default, the GINI) for the best possible split is less than 1% of the impurity in the root node. Also, splits are usually only attempted on nodes with at least 5% of the number of total observations. This allows for somewhat larger trees in case-control studies with relatively few observations. Using those parameters, we grow the trees to full size without model selection and pruning. In our simulations, this provided some additional computational benefit as it was not necessary to carry out cross-validation, without compromising the quality of the imputations. By default, we iterate 10 times through the set of missing variables ("sweeps" through the data) before imputing the missing values. However in data sets with severe missingness, more sweeps might be beneficial.

Multiple Imputation

The uncertainty of imputations is addressed by multiple imputation [Little and Rubin, 1987; Schafer, 1997]. Multiple imputation is a Monte Carlo technique which draws multiple samples from the probability distribution of predicted missing values. As described previously, we draw 10 samples from the resulting joint distribution of missing data at convergence, whether it is from EM, WEM or tree algorithm. Each imputed sample is analyzed by standard methods, and the results are combined across 10 samples to get parameter estimates and their standard errors. The details of multiple imputation have been documented in Little and Rubin [1987] and Schafer [1997].

Simulations

Our simulation studies involved drawing case-control samples from a population, randomly masking a proportion of SNPs as missing, and imputing them by the methods under investigation. We adopted an eight-haplotype distribution based on four SNPs in the progesterone receptor (PGR) gene [Kraft et al., 2005]. Previously, De Vivo et al. [2002] found that a

G/A polymorphism in the PGR gene may be associated with an increased risk of endometrial cancer. Kraft et al. [2005] genotyped four haplotype tagging SNPs (htSNPs) in case control data in order to compare several methods currently used in haplotype-disease association studies. Table I shows the distribution of eight haplotypes estimated in Kraft et al. [2005]. Based on these frequencies and assuming Hardy-Weinberg equilibrium, we created a population of 100,000 individuals with diploid genotypes.

Table I. PGR haplotype frequencies (Kraft et al. 2005) used in the simulation study. Haplotype 1000 is associated with the disease outcome.

Haplotype	Frequency
0000	0.3265
0001	0.1327
0100	0.0306
0101	0.0408
1000	**0.1613**
1010	0.0408
1100	0.0204
1110	0.2449

We added a disease-association signal to haplotype **1000** through a logistic penetrance function

$$\text{logit}(\Pr(\mathbf{D} = 1|\mathbf{H})) = -3 + \beta \cdot (\text{number of copies of } h_{1000}), \qquad (4)$$

with $\beta = 0$, 1, or 2. **D** is the dichotomous disease status, and **H** refers to the haplotype pair for an individual. We randomly sampled 100 cases and 300 controls from the population. Either 10% or 20% of the SNPs were made missing completely at random. These missing SNPs were imputed ten times using the EM, WEM, and tree approach. To construct a baseline for the imputation comparison, we used the observed marginal SNP genotype distribution to impute the missing ones. We call this method the "naive" approach, as it uses no information of other SNPs or the response. We calculated the imputation error probability for each SNP using a 0/1 loss function. That is, we coded each genotype as 0 (homozygous wide type), 1 (heterozygotes), and 2 (homozygous mutant) and any difference in imputed genotype was counted as an error. We explored a variety of other error functions, reaching similar conclusions about the various approaches.

While we predisposed disease risk on the haplotype level, we analyzed the imputed data using SNP based logistic regression models. We first considered marginal SNP association with disease by fitting a logistic regression model of the form

$$\text{logit}(\Pr(\mathbf{D}|\text{SNPs})) = \alpha_0 + \alpha_1 x, \tag{5}$$

where x denotes the number of variant alleles (0,1,2) for a particular SNP. For simplicity we treated x as a continuous variable so that having two copies doubles the effect of having one copy. We also investigated the effect of imputation on interactions using the model

$$\text{logit}(\Pr(\mathbf{D}|\text{SNPs})) = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 x_1 x_2. \tag{6}$$

Similarly x_1 and x_2 are the coding variables for SNP 1 and SNP 2, respectively. γ_3 is the interaction parameter of interest. We compared the parameter estimates using various approaches of imputing SNP data with the true values, that can be computed by fitting (5) and (6) to the whole population.

Data application

We used a recently published case-control dataset on developmental dyslexia (DD) to compare various imputation approaches. Cope et al. [2005] performed a high-density linkage disequilibrium screen in a 575-kb region of chromosome 6p22.2 with both case-control and family data. We used the case-control data for imputation, which includes 248 cases and 273 controls. We only used the six SNPs in the *KIAA0319* gene, since there is strong evidence that it is a susceptibility gene, and pair-wise LD analysis suggests that these 6 SNPs are located in a block so that they are good candidates for a haplotype reconstruction. Note these 6 SNPs are all htSNPs and inter-SNPs correlations are weak according to Table 2 in Cope et al. [2005]. Table II shows the number of missing values for each of these SNPs, separately for cases and controls. 25 probands from the later family study were also included as cases, who do not have genotypes for SNP *rs6911855* and *rs6939068*. This is because these two SNPs are not significant in the initial case-control screen, hence not genotyped in the later family study. Cope et al. [2005] ignored missing data and analyzed the data in a SNP by SNP fashion.

Table II. Percentage of missing SNPs in the case-control study in Cope et al. (2005).

	Case ($n = 248$)		Control ($n = 273$)	
SNP	#	%	#	%
rs4504469	8	3.2	9	3.3
rs6911855	30	12.1	8	2.9
rs6939068	48	19.4	25	9.2
rs2179515	16	6.5	16	5.9
rs6935076	17	6.9	18	6.6
rs2038137	19	7.8	16	5.9
Total	138	9.3	92	5.6

Note - *rs6939068* and *rs6939068* have extra missing values in cases since 25 probands from the later family study are included. By design these 25 cases do not have the genotypes for *rs6939068* and *rs6939068*.

We reanalyzed the marginal SNP association in DD data via the multiple imputation approaches under investigation. The SNP-disease association was modeled as in (5). To evaluate the imputation errors, we randomly generated extra missing values and computed the probability of false imputation for the additional missing data. Parameter estimates for model (5) and (6) were computed with the extra missing data imputed, but the original missing data from Table II were left unimputed. The "true values" of parameter estimates are therefore computed from the original data with missing values. We compared the bias and sampling variance as in the simulation study.

Results

For this relatively simple LD block with 4 SNPs in the simulation study, both the EM and WEM approach yield better predictions of the missing SNPs than the tree approach (Table III), while all three approaches show a marked improvement over the naive approach. When there is no association between the SNPs and the outcome ($\beta = 0$) the EM and WEM approaches performed equally well and the tree approach makes roughly 2–3% more errors. However, when there is a disease risk associated with haplotype **1000** the WEM approach yields more accurate imputations than the EM approach. As β increases to 2, the advantage of WEM for SNP1 imputation becomes substantial: WEM produces almost 5–6% less errors than EM. This was expected since the case-control status influences the estimation of haplotype frequencies. When the association is absent ($\beta = 0$) or small ($\beta = 1$) the tree approach performs comparably to the EM and WEM approach. Given that the tree algorithm treats SNPs as 0/1/2 categorical variables, thus completely ignores the underlying haplotype structure, such performance is impressive. When the association is strong ($\beta = 2$), the tree approach even outperforms EM for SNP1, presumably because the strength of the association now overcomes the incorrect model. A graphical representation of Table III as well as the following tables can be found as supplementary materials at `http://biostat.jhsph.edu/∼iruczins/supplements/05.comparison`.

Table III. Mean imputation errors in the simulated data of four SNPs on the PGR gene for four imputation approaches.

		10% missing data				20% missing data			
	Approach	SNP1	SNP2	SNP3	SNP4	SNP1	SNP2	SNP3	SNP4
$\beta = 0$									
	Naive[a]	0.625	0.596	0.568	0.449	0.625	0.595	0.567	0.449
	EM	0.412	0.390	0.243	0.379	0.427	0.407	0.271	0.385
	WEM	0.412	0.390	0.243	0.379	0.427	0.406	0.271	0.385
	Tree	0.440	0.397	0.260	0.399	0.461	0.411	0.292	0.415
$\beta = 1$									
	Naive	0.627	0.589	0.560	0.441	0.627	0.589	0.560	0.441
	EM	0.433	0.383	0.245	0.369	0.448	0.399	0.273	0.375
	WEM	0.415	0.381	0.241	0.369	0.431	0.396	0.269	0.375
	Tree	0.449	0.389	0.263	0.389	0.471	0.407	0.296	0.403
$\beta = 2$									
	Naive	0.628	0.587	0.557	0.438	0.627	0.588	0.557	0.438
	EM	0.443	0.380	0.246	0.365	0.457	0.397	0.273	0.371
	WEM	0.386	0.375	0.233	0.363	0.402	0.391	0.257	0.370
	Tree	0.422	0.388	0.262	0.385	0.443	0.398	0.292	0.399

Note - Each number is the average of imputation error probabilities from 5000 simulations.
[a] this method imputes the missing by the marginal distribution of available SNP genotypes.

In Table IV we compare the effects of the different imputation approaches on estimating the log-odds ratio for SNP1, as modeled by α_1 in (5). In this table, the lines "True data" refer to the case-control data before SNPs were made missing, as these are the best imputations one can obtain. In all scenarios, WEM has the smallest bias, whereas EM has the smallest sample variance. When there is no SNP disease association, all three approaches perform comparably in square root of mean square error (RMSE). When haplotype **1000** is associated with the disease outcome, the WEM approach has less bias than the tree approach, while the EM approach has more bias. As a result, the WEM approach has the smallest RMSE among all methods when there is a strong signal. Note that the WEM approach consistently outperforms the complete-case approach, suggesting that there is something to be gained by imputing SNPs. The tree approach has a relatively large variance, but it has less bias than the EM approach, and it outperforms EM in terms of RMSE when $\beta = 2$. The large variance reflects the inherent variability of the tree-based regression, as trees are known to be unstable predictors [Hastie et al., 2002].

We further investigated the effect of imputation on the estimation of interaction parameters. For SNP interactions, the complete-case analysis hurts more as 10% missing values in each SNP may result in up to 20% missing in either one of two SNPs. This is confirmed by the results in Table V. All imputation approaches substantially reduce the sampling variance and RMSE of $\widehat{\gamma}_3$, in comparison to a complete-case analysis. Because of the severity of missingness and the higher variability in estimating interactions, sampling variance dominates bias in RMSE calculation so that the EM approach generates the smallest RMSE. Interestingly, the EM and the tree approach produce a smaller sampling variance and RMSE than the true data. A similar pattern is present when we compare RMSE for the SNP1 main effect at $\beta = 0$ in Table IV. Further inspection of the simulation results suggests that the imputation of missing data always shrinks the parameter estimates slightly toward null, since no imputation will ever achieve 100% accuracy. This shrinkage effect may lower the sampling variance, and thus generate a smaller RMSE than using the true data.

Cope et al. [2005] estimated the odds ratios for the allele effect for each SNP ignoring missing data. They concluded that four of six SNPs in *KIAA0319* are significantly associated with DD at a significance level of 0.05. We carried out a multiple imputation analysis to see

Table IV. The effect of different imputation approaches on the marginal association parameter α_1 for SNP1.

	10% missing data			20% missing data		
Approach	Bias	SD($\widehat{\alpha}_1$)	RMSE	Bias	SD($\widehat{\alpha}_1$)	RMSE
β=0						
True data	−0.011	0.168	0.168	−0.011	0.168	0.168
Complete-case	−0.013	0.176	0.177	−0.011	0.187	0.187
EM imputed data	−0.014	0.163	0.164	−0.016	0.160	0.160
WEM imputed data	−0.011	0.173	0.173	−0.009	0.178	0.178
Tree imputed data	−0.013	0.166	0.166	−0.014	0.163	0.164
β=1						
True data	0.006	0.163	0.163	0.006	0.163	0.163
Complete-case	0.006	0.173	0.173	0.005	0.182	0.181
EM imputed data	−0.051	0.162	0.169	−0.107	0.155	0.188
WEM imputed data	−0.001	0.174	0.174	−0.005	0.181	0.181
Tree imputed data	−0.020	0.173	0.174	−0.049	0.183	0.189
β=2						
True data	0.012	0.181	0.181	0.012	0.181	0.181
Complete-case	0.011	0.190	0.190	0.012	0.204	0.204
EM imputed data	−0.104	0.169	0.198	−0.211	0.162	0.266
WEM imputed data	0.003	0.184	0.184	−0.003	0.193	0.193
Tree imputed data	0.005	0.190	0.190	−0.011	0.211	0.211

Note - For each combination of the missing proportion and the imputation approach, the logistic model logit(Pr($\mathbf{D}$|SNPs)) = $\alpha_0 + \alpha_1 x$ was fitted; x is the continuous coding variables for SNP1, valued at 0, 1,and 2. The true value of the parameter is obtained by fitting the model to the population. Bias is the mean difference between estimated parameters and the true value. SD($\widehat{\alpha}_1$) is the sample standard deviation of the estimated parameters. RMSE is the square root of the mean square error.

whether imputation of missing data changes these conclusions. Six individuals with all SNPs missing were left out of this analysis. We verified that the SNP effect is indeed additive, and applied the univariate logistic regression as in (5) to each SNP. Table VI compares the log-odds ratio estimates and significant levels. We explain the results by dividing six SNPs into two groups. Group 1 contains *rs4504469*, *rs2179515*, *rs6935076* and *rs2038137*, each of which has less missing values roughly balanced between cases and controls (Table II). Group 2 contains *rs6939068* and *rs6911855*. Both SNPs have more missing values in total, and more missing values in cases than controls (Table II). For SNPs in group 1 the imputation has little effect on the standard errors. The effect of the sample size increase after imputation seems to be canceled by the extra variability raised by multiple imputation. This is perhaps because the rate of missing values is small for these loci and LD is weak (as suggested by Table 2

Table V. The effect of different imputation approaches on the interaction parameter γ_3 for SNP1 and SNP2.

	10% missing data			20% missing data		
Approach	Bias	SD($\widehat{\gamma}_3$)	RMSE	Bias	SD($\widehat{\gamma}_3$)	RMSE
β=0						
True data[a]	−0.022	0.255	0.255	−0.022	0.255	0.255
Complete-case	−0.029	0.281	0.282	−0.022	0.319	0.320
EM imputed data	−0.022	0.242	0.243	−0.017	0.234	0.234
WEM imputed data	−0.026	0.262	0.263	−0.024	0.271	0.272
Tree imputed data	−0.020	0.238	0.239	−0.015	0.231	0.231
β=1						
True data	0.010	0.303	0.303	0.010	0.303	0.303
Complete-case	0.010	0.332	0.332	0.000	0.366	0.366
EM imputed data	0.051	0.282	0.287	0.097	0.249	0.267
WEM imputed data	0.036	0.306	0.308	0.058	0.310	0.315
Tree imputed data	0.040	0.291	0.294	0.069	0.268	0.277
β=2						
True data	−0.041	0.368	0.370	−0.041	0.368	0.370
Complete-case	−0.049	0.425	0.427	−0.049	0.467	0.470
EM imputed data	0.060	0.333	0.338	0.145	0.281	0.316
WEM imputed data	0.002	0.394	0.394	0.054	0.400	0.403
Tree imputed data	0.018	0.366	0.366	0.084	0.327	0.337

Note - For each combination of the missing proportion and the imputation approach, the logistic model logit(Pr($\mathbf{D}$|SNPs)) = $\gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 x_1 x_2$ was fitted; x_1 and x_2 are the continuous coding variables for SNP1 and SNP2, valued at 0, 1,and 2. The true value of the parameter is obtained by fitting the model to the population. Bias is the mean difference between estimated γ_3 and the true value. SD($\widehat{\gamma_3}$) is the sample standard deviation of the estimated γ_3. RMSE is the square root of the mean square error.

[a] original data without any missing values.

in Cope et al. [2005]). In comparison to the complete-case analysis, the point estimates of log-OR using the various imputation approaches are shrunken toward null, especially for the EM and tree approach. For SNPs in group 2, however, standard errors become smaller and point estimates are enlarged by both the EM and WEM imputations. Hence the resulting p-values are smaller than the complete-case analysis, especially for SNP *rs6911855*. The difference of group 1 and 2 is driven by the 25 proband-cases, who have two SNPs in group 2 missing. The WEM approach seems to capture the missing pattern depending on the disease status, therefore yield more significant results than the other approaches.

To compare the accuracy of three imputation approaches, we again randomly removed an extra 5%, 10%, and 15% of the SNPs from the dataset of Cope et al. [2005]. Table

VII shows the comparison of imputation error probabilities for the additional missing SNPs stratified by SNP, imputation approach, and missing percentage. Similar to Table III, all three approaches work much better than the naive approach. The WEM approach performs consistently better than the EM and tree approaches, although the improvement of WEM over EM is for most cases less than 1%. This may be explained by the weak LD among the 6 SNPs, since haplotype ambiguity is so substantial that knowing case-control status does not gain much in imputation. On the other hand, the accuracy of the tree approach is only 1-2% lower than two haplotype-based approaches, suggesting that in practice the tree approach may be sufficiently accurate to characterize the inter-SNP correlation in a modest LD block.

Table VIII shows the biases, sample standard deviations and RMSE for two SNPs found to be most significantly associated with DD in Cope et al. [2005], using different imputation approaches. These statistics are conditional on the original data from Cope et al. [2005]. That is, the "true values" of the parameters are obtained from the original data with the original missing values and no imputation. Likewise, the "Complete-case" here refers to datasets with both the original missing SNPs and extra missing data removed, again serving as the baseline for comparison. The first six columns compare the estimates of SNP marginal effects. It appears that all three imputation approaches improve the SD and RMSE over the complete-case analysis. Among them, the WEM approach is effectively unbiased and it has the smallest RMSE under almost all conditions. Interestingly, the tree approach achieves the second best performance in bias reduction and RMSE, superior to the EM approach in most situations. This seems contradictory to the comparison in Table VII, where the tree approach makes more imputation errors than the EM approach. Since we count 1 error whenever the imputed SNP genotype is different from the true genotype, the impact of different genotype errors on association parameter estimates may be different. For example, imputing a missing SNP with true genotype "2" to be "1" has less effect than imputing it to be "0". It is possible that the impact of imputation errors on estimating association parameters is smaller in the tree approach than that in the EM approach, since the former use case-control status in the imputation.

Cope et al. [2005] found a significant interaction between *rs4504469* and *rs6935076*. In the last three columns in Table VIII, we compared the effects of imputation on the interaction parameter in the logistic regression model (6). Similarly to what we found in Table V, all imputation approaches show a marked improvement on SD and RMSE compared to no imputation at all (complete-case). EM has the smallest RMSE owing to its low variability, even though WEM has the smallest bias.

Discussion

Despite the fact that missing SNPs are quite common in genetic association studies, the impact of imputation on SNP association inference has not been adequately studied. In this article, we developed and compared the haplotype-based and the tree-based imputation approaches in case-control data. Our results suggest that in general there is benefit from imputation over the commonly used complete-case analysis. As we expected, the benefit of imputation is greater in estimating interaction parameters than that in estimating marginal parameters (Tables V and VIII). Haplotype-based approaches show slightly better imputation accuracy and better inference properties than the nonparametric tree based approach, when LD blocks are present. The performance of the tree approach is rather impressive given the fact that it ignores the underlying haplotype structure. The weighted EM approach, not surprisingly, is superior to the EM approach if there is a SNP-disease association.

Imputing missing SNPs usually helps association inference in increasing the efficiency without adding noticeable bias, yet at the price of some extra variability from the uncertainty in the imputation. With LD structure existing between SNPs, the added sample size usually outweighs the imputation uncertainty and the standard error decreases. This is seen in our simulation study (Tables IV and V) and the data application (Table VIII). The advantage of imputation can be substantial when a regression model with multiple SNPs is employed (Tables V and VIII). In some cases where missingness rate is low and LD is weak, the imputation may not help to gain efficiency for marginal parameters (Table VI). On the other hand, imputation of the missing SNPs could also help to correct bias. In our data application, the fractions of missing values for SNP *rs6911855* and SNP *rs6939068* differ between cases and controls. The parameter estimates and association inferences for these two SNPs were changed greatly by multiple imputation using the WEM approach (Table VI), suggesting complete-case analysis may cause bias in this scenario. Our overall assessment is that performing multiple imputation up-front yields better inferences than the complete-case analysis in SNP association studies, particularly when regression models with multiple SNPs are involved.

Haplotype analysis is becoming increasingly popular in genetic association studies, whereas tree-based approaches start to draw attention in studies with a large number of SNPs. To our knowledge, this is the first paper directly comparing haplotype and tree based approaches. The imputation accuracy serves as an indicator as how well the inter-SNPs correlation structure is captured by nonparametric tree regression. Evidently, the tree approach produces slightly more imputation errors than the haplotype approaches and it is more variable by nature. However its advantages are apparent: it is computationally efficient, it easily accommodates disease status, extra covariates, and a large number of SNPs. In some situations with weak LD blocks present (marginal effects in Table VIII), the tree approach even outperforms the EM approach in both bias reduction and MSE. In many genetic epidemiological studies subjects complete a questionnaire, which may contain dozens of relevant environmental and demographic variables. The tree algorithm can handle an arbitrary number of these variables, as the splits in the decision trees are completely data driven. Computing time so far has never been an issue in our analyses (typical data we see have up to a few thousand observations and a few hundred variables). Considering the increasing number of genome-wide SNP association studies carried out today, we believe that the tree approach provides a competitive alternative for the imputation of missing SNP values.

For the benefit of imputation, the information of disease status is secondary to the correlation between adjacent SNPs in the data we studied. That is perhaps why the improvement of the WEM approach over the EM approach is much smaller compared to that of the EM approach over naive imputation. In virtually every situation we examined, the WEM approach produces minimal bias since it uses the critical information of case-control status. On the other hand, the WEM approach incurs more variability than the EM approach since it involves more parameters in the modeling. That said, we recognize that sometimes one may be willing to take the unbiased approach even it has a larger variance. Since we generally do not know whether there is association between haplotype and disease beforehand, using the WEM approach to impute the missing data seems a “safe” choice.

There are other algorithms to reconstruct haplotypes, and therefore impute the missing SNPs. For example, PHASE employs a MCMC approach to infer the haplotypes from unphased genotypes, using priors based on coalescent theory and taking account of the decay of linkage disequilibrium [Stephens et al., 2001; Stephens and Scheet 2003]. However, PHASE, as well as other Baysian approaches, is designed for inferring haplotype in a population and thus does not incorporate the case-control status to estimate the haplotype frequencies. This may introduce bias to association inference after imputing the missing SNPs, similarly to the EM algorithm. We tried PHASE to impute the missing SNPS in the developmental dyslexia data. The imputation accuracy was about the same as the EM algorithm because of the weak LD structure between the 6 tagSNPs, yet the computing time was significantly longer than the other three approaches. For the convenience of computation and implementation we did not include PHASE in our comparisons.

Acknowledgments

We want to thank Dr. Peter Holmans and the other members of the developmental dyslexia (DD) study for allowing us to use the case control data. This research was supported in part by NIH grants CA 74841 (JYD, IR, MLB, CK), CA 105069 (IR), CA 53996 (MLB, CK), CA 90998 (MLB), and HL 74745 (CK).

References

Barnby G, Abbott A, Sykes N, Morris A, Weeks DE, Mott R, Lamb J, Bailey AJ, Monaco AP and the International Molecular Genetics Study of Autism Consortium (IMGSAC). 2005. Candidate-gene screening and association analysis at the autism-susceptibility locus on chromosome 16p: evidence of association at GRIN2A and ABAT. Am J Hum Genet 76:950-966.

Botstein D, Risch N. 2003. Discovering genotypes underlying human phenotypes: past success for Mendelian disease, future approaches for complex disease. Nat Genet Suppl 33:228-237.

Breiman L, Friedman J, Olshen R, Stone C. 1984. Classification and regression trees. Belmont, CA: Wadsworth International Group.

Brewster A, Jorgensen T, Ruczinski I, Huang H, Hoffman S, Thuita L, Newschaffer C, et al. 2005. Polymorphisms of the DNA repair genes XPD (Lys751Gln) and XRCC1 (Arg399Gln and Arg194Trp): Relationship to Breast Cancer Risk and Familial Predisposition to breast cancer. Breast Cancer Research and Treatment (in press).

Bureau A, Dupuis J, Falls K, Lunetta K, Hayward B, Keith T, Van E. 2005. Identifying SNPs predictive of phenotype using random forests. Genet Epidemiol 28:171-182.

Cope N, Harold D, Hill G, Moskvina V, Stevenson J, Holmans P, Owen M, et al. 2005. Strong evidence that KIAA0319 on chromosome 6p is a susceptibility gene for developmental dyslexia. Am J Hum Genet 76:581-591.

Cordell H, Clayton D. 2002. 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 70:124-141.

De Vivo I, Huggins G, Hankinson S, Lescault P, Boezen M, Colditz G, Hunter D. 2002. A functional polymorphism in the promoter of the progesterone receptor gene associated with endometrial cancer risk. Proc Natl Acad Sci USA 99:12263-12268.

Epstein M, Satten G. 2003. Inference on haplotype effects in case-control studies using unphased genotype data. Am J Hum Genet 73:1316-1329.

Excoffier L, Slatkin M. 1995. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Bio Evol 12:921-927.

Greenland S, Finkle W. 1995. A critical look at methods for handling missing covariates in epidemiologic regression analysis. Am J Epidemiol 142:1255-1264.

Harrell F. 2001. Regression modelling strategies. New York: Springer.

Hastie T, Tibshirani R, Friedman J. 2001. The elements of statistical learning. New York: Springer.

Hu N, Wang C, Hu Y, Yang H, Giffen C, Tang Z, Han X, et al. 2005. Genome-wide association study in esophageal cancer using GeneChip Mapping 10K Assay. Cancer Res 65:2542-2546.

Kraft P, Cox D, Paynter R, Hunter F, De Vivo I. 2005. Accounting for haplotype uncertainty in matched association studies: a comparison of simple and flexible techniques. Genet Epidemiol 28:261-272.

Lake S, Lyon H, Tantisira K, Silverman E, Weiss S, Laird N, Schaid D. 2003. Estimation and tests of haplotype-environmental interaction when linkage phase is ambiguous. Hum Hered 55:56-65.

Lin S, Cutler D, Zwick M, Chakravarti A. 2002. Haplotype inference in random population samples. Am J Hum Genet 71:1129-1137.

Little R, Rubin D. 1987. Statistical analysis with missing data. New York: John Wiley & Sons.

Niu T, Qin Z, Xu X, Liu J. 2002. Bayesian haplotype inference for multiple linked single nucleotide polymorphisms. Am J Hum Genet 70:157-169.

Risch N. 1990. Linkage strategies for genetically complex traits. I. Multi-locus models. Am J Hum Genet 46:222-228.

Risch N. 2000. Searching for genetic determinants in the new millennium. Nature 405:847-856.

Ruczinski I, Kooperberg C, LeBlanc M. 2003. Logic regression. J Comput Graph Stat 12:475-511.

Schafer J. 1997. Analysis of incomplete multivariate data. London: Chapman & Hall.

Schaid D, Rowland C, Tines D, Jacobson R, Poland G. 2002. Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 70:425-434.

Stephen M, Scheet P. 2003. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am J Hum Genet 76:449-462.

Stram D, Leigh PC, Bretsky P, Freedman M, Hirschhorn J, Altshuler D, Kolonel L, et al. 2003. Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. Hum Hered 55:179-190.

Tanner M, Wong W. 1997. The calculation of posterior distributions by data augmentation. JASA 82:528-550.

Therneau T, Atkinson E. 1997. An introduction to recursive partitioning using the RPART routines. Technical Report Series No.61. Department of Health Science Research, Mayo Clinic, Rochester, Minnesota.

Zhang H, Bonney G. 2000. Use of classification trees for association studies. Genet Epidemiol 19:323-332.

Zhao L, Li S, Khalid N. 2003. A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. Am J Hum Genet 72:1231-1250.

Table VI. A comparison of the log-odds ratio estimates by different imputation methods for the developmental dyslexia data.

Approach	Complete-case			EM imputed data			WEM imputed data			Tree imputed data		
SNP	log-OR	SE	P-val	log-OR	SE	P-val	log-OR	SE	P-val	log-OR	SE	P-val
rs4504469	−0.417	0.018	0.002	−0.414	0.018	0.002	−0.426	0.018	0.001	−0.416	0.018	0.002
rs6911855	0.658	0.138	0.076	0.684	0.135	0.063	0.720	0.138	0.052	0.614	0.138	0.098
rs6939068	0.637	0.123	0.070	0.648	0.111	0.052	0.700	0.114	0.038	0.584	0.119	0.090
rs2179515	−0.362	0.019	0.008	−0.357	0.018	0.008	−0.357	0.019	0.009	−0.351	0.019	0.010
rs6935076	0.396	0.019	0.005	0.378	0.019	0.006	0.380	0.019	0.006	0.356	0.020	0.012
rs2038137	−0.439	0.019	0.001	−0.411	0.018	0.002	−0.422	0.018	0.002	−0.412	0.019	0.003

Note - the estimates are based on 10 imputations.

Table VII. The comparison of imputation error probabilities for the developmental dyslexia data by four methods.

	Approach	rs4504469	rs6911855	rs6939068	rs2179515	rs6935076	rs2038137	Average
5% missing	Naive	0.609	0.117	0.133	0.595	0.596	0.600	0.447
	EM	0.370	0.034	0.033	0.089	0.446	0.092	0.181
	WEM	0.367	0.032	0.032	0.085	0.442	0.091	0.178
	Tree	0.379	0.038	0.036	0.106	0.456	0.114	0.192
10% missing	Naive	0.609	0.114	0.137	0.594	0.597	0.600	0.447
	EM	0.376	0.039	0.039	0.098	0.447	0.104	0.187
	WEM	0.373	0.037	0.039	0.095	0.442	0.103	0.185
	Tree	0.388	0.041	0.041	0.127	0.462	0.135	0.202
15% missing	Naive	0.610	0.114	0.136	0.594	0.595	0.600	0.447
	EM	0.380	0.042	0.046	0.110	0.451	0.115	0.194
	WEM	0.377	0.042	0.045	0.107	0.446	0.114	0.191
	Tree	0.396	0.044	0.047	0.147	0.466	0.151	0.212

Note - the numbers are the averages of imputation error probabilities from 200 simulations.

Table VIII. The comparisons of the marginal log-odds ratio estimates for SNPs *rs4504469* , *rs6935076* and their interactions.

	rs4504469[a]			*rs6935076*[a]			interaction[b]		
Approach	Bias	SD($\widehat{\alpha}_1$)	RMSE	Bias	SD($\widehat{\alpha}_1$)	RMSE	Bias	SD($\widehat{\gamma}_3$)	RMSE
5% missing									
Complete-case	0.0024	0.0291	0.0291	−0.0021	0.0306	0.0306	−0.0010	0.0701	0.0699
EM imputed data	0.0099	0.0214	0.0236	−0.0109	0.0245	0.0267	−0.0137	0.0536	0.0552
WEM imputed data	0.0043	0.0220	0.0224	−0.0010	0.0256	0.0255	−0.0012	0.0570	0.0568
Tree imputed data	0.0095	0.0213	0.0233	−0.0067	0.0257	0.0265	−0.0215	0.0547	0.0587
10% missing									
Complete-case	0.0034	0.0438	0.0438	−0.0035	0.0479	0.0479	−0.0073	0.1047	0.1047
EM imputed data	0.0197	0.0303	0.0360	−0.0220	0.0340	0.0404	−0.0302	0.0713	0.0772
WEM imputed data	0.0089	0.0336	0.0347	−0.0018	0.0374	0.0373	−0.0049	0.0802	0.0802
Tree imputed data	0.0179	0.0305	0.0353	−0.0148	0.0348	0.0378	−0.0467	0.0787	0.0914
15% missing									
Complete-case	0.0011	0.0557	0.0555	−0.0001	0.0623	0.0621	−0.0069	0.1299	0.1298
EM imputed data	0.0271	0.0356	0.0447	−0.0321	0.0434	0.0539	−0.0412	0.0819	0.0915
WEM imputed data	0.0085	0.0421	0.0429	0.0020	0.0518	0.0517	−0.0063	0.0925	0.0925
Tree imputed data	0.0258	0.0368	0.0448	−0.0187	0.0468	0.0503	−0.0684	0.0869	0.1104

[a] Marginal effect: logistic model logit(Pr($\mathbf{D}$|SNPs)) $= \alpha_0 + \alpha_1 x$ was fitted; x is the continuous coding variables for the targeted SNP. The true value of the parameter is obtained by fitting the model to the original developmental dyslexia data; Bias, SD, and RMSE are computed from 200 iterations.

[b] Interaction: logistic model logit(Pr($\mathbf{D}$|SNPs)) $= \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 x_1 x_2$ was fitted. x_1 and x_2 are the 0/1/2 continuous coding variables for *rs4504469* and *rs6935076*.