Assessing measurement error correction for genotype imputation in GWAS

Quenna Wong

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2012

Reading Committee:

Robyn McClelland, Chair

N. David Yanez, Chair

Program Authorized to Offer Degree:
UW Biostatistics

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

The author wishes to express sincere appreciation to her advisors and thesis committee co-chairs, Robyn McClelland and David Yanez, for their guidance, support, and of course, patience.

The author also wishes to thank the investigators, the staff, and the participants of the MESA study for their valuable contributions. MESA and the MESA SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts N01-HC-95159 through N01-HC-95169 and UL1-RR-024156. Funding for genotyping was provided by NHLBI Contract N02-HL-6-4278 and N01-HC-65226. A full list of participating MESA investigators and institutions can be found at http://www.mesa-nhlbi.org.

Chapter 1

INTRODUCTION

A common situation arises in the analysis of quantitative data when measured variables cannot be precisely ascertained. The "true" values of the variables may be unattainable or cost prohibitive to acquire. In such cases, the observed variable or surrogate that is used in place of the unobserved variable is said to be measured with error.

1.1 Examples of measurement error in medical research

In cardiovascular research, intimal-media thickness (IMT) measurements of the common carotid artery wall are obtained using ultrasound scans. These data are subject to measurement error due to sonographers and readers of the sonographic images. Left-ventricular mass (LV mass) measurements are obtained by magnetic resonance imaging (MRI). Coronary artery calcium scores (CAC) measure the amount of calcium in the coronary arteries and are obtained using computed tomography (CT) imaging. All of these imaging modalities are commonly used variables, and are subject to measurement error.

Cardiovascular disease risk factors are also affected by measurement error, including blood pressures, serum cholesterol, and glucose levels. Blood pressure readings are subject to temporal, instrumental, and technician variation. Serum measures are subject to laboratory and assay variation. Other examples where measurement error arises include dietary consumption data (e.g., food frequency questionnaires, interviews in nutrition studies), absorbed herbicide volumes (in agricultural studies), and NO_2 exposure or forced expiratory volume (FEV) measurements in studies of

pulmonary function.

Advancements in technology coupled with new computational methods have allowed disease research to expand into an era involving analysis of genetic data, measurements of many millions of variants. Such research often aims to identify genetic information underlying disease and risk factors for disease. Nonetheless, it is not surprising that acquisition of these millions of pieces of genetic information is also subject to measurement error. This thesis will focus on measurement error correction techniques geared towards genetic data. The following section gives an overview of this type of data.

1.1.1 Genetics and GWAS

It is estimated that unrelated humans have roughly 99% of their DNA in common from one person to the next, and that the differences in the remaining 1% of DNA play a role in determining the unique characteristics of an individual including differential health and disease susceptibility. These genetic variants span a range of complexity from single positional substitutions to insertions, deletions, repetitions, and inversions which may involve larger sections of DNA.

In a typical genome-wide association study (GWAS), researchers try to pinpoint specific Single Nucleotide Polymorphisms (SNPs, a variation of a nucleotide at any position in a DNA sequence) associated with a particular trait, or phenotype of interest. For each trait, this results in performing one statistical test of association for each SNP studied − in other words, often a million tests. As a result, investigators seek out collaborations with other studies that have similar phenotypes in order to increase power and minimize detection of false signals. However, the set of genotyped SNPs is usually not identical across the studies. To allow the studies to be combined, imputation is necessary in order to estimate genotypes for SNPs missing in one study but not another. Imputation is performed for each study to generate a common and more expansive set of millions of SNPs. The imputation process is described in greater

detail in Chapter 3, outlining how this process generates data that may be considered as the true genotype measured with error.

1.2 "The double whammy" - bias and loss of power

If scientific interest in a study is to investigate an association of an outcome with the *true* unobserved predictor variable (as in a GWAS), it is likely the estimated association will be biased, possibly severely, because of measurement error. In addition, standard methods of statistical analyses (for example, ordinary least squares regression) can yield reduced power to detect associations and may produce misleading graphical presentations (Carroll et al., 2006, pp. 1-2, 18-20).

1.2.1 Bias

In regression analysis, measurement error bias may occur when one estimates the associations between an outcome and observed mismeasured predictor variables. Measurement error bias generally does not occur if one solely has imprecisely measured outcome variables. When measurement error is present in predictor variables, it is frequently assumed that there is a "bias toward the null" and the measurement error is consequently ignored. Unfortunately, this bias to the null is not always true. The direction and magnitude of the bias are influenced by the relationship between the observed variable and the unobserved variable as well as the relationship between the observed variable and the other predictor variables in the regression model.

To illustrate the bias in the simplest case, consider a simulation using simple linear regression (Figure 1.1). In this example, we simulate 30 true (unobserved) values of the predictor of interest, x, as independent observations from a standard normal distribution. Our regression model is $y = 0 + 1 \cdot x + \epsilon$, where $x \sim N(0, 1)$ and model error $\epsilon \sim N(0, 0.1)$. Next, we simulated 30 independent measurement errors, u, from a standard normal distribution to compute the observed imprecisely measured predictors, $w = x + u$.

The expected coefficient estimate of the predictor, x, is 1.0, for the regression of y on x. For the 30 simulated (x, y) pairs, we obtain an estimate of 1.02 (SE=0.05). Modeling y on the observed predictor, w, we obtain a coefficient estimate of 0.53 (SE=0.07). The horizontal grey lines in Figure 1.1 show the observed (open circles) and unobserved (filled circles) values, illustrating the horizontal spreading in the observed regression line and attenuation of the coefficient estimate of the observed predictor. The figure and estimated standard errors for the two models reveal that the observed data yield more variable estimates.

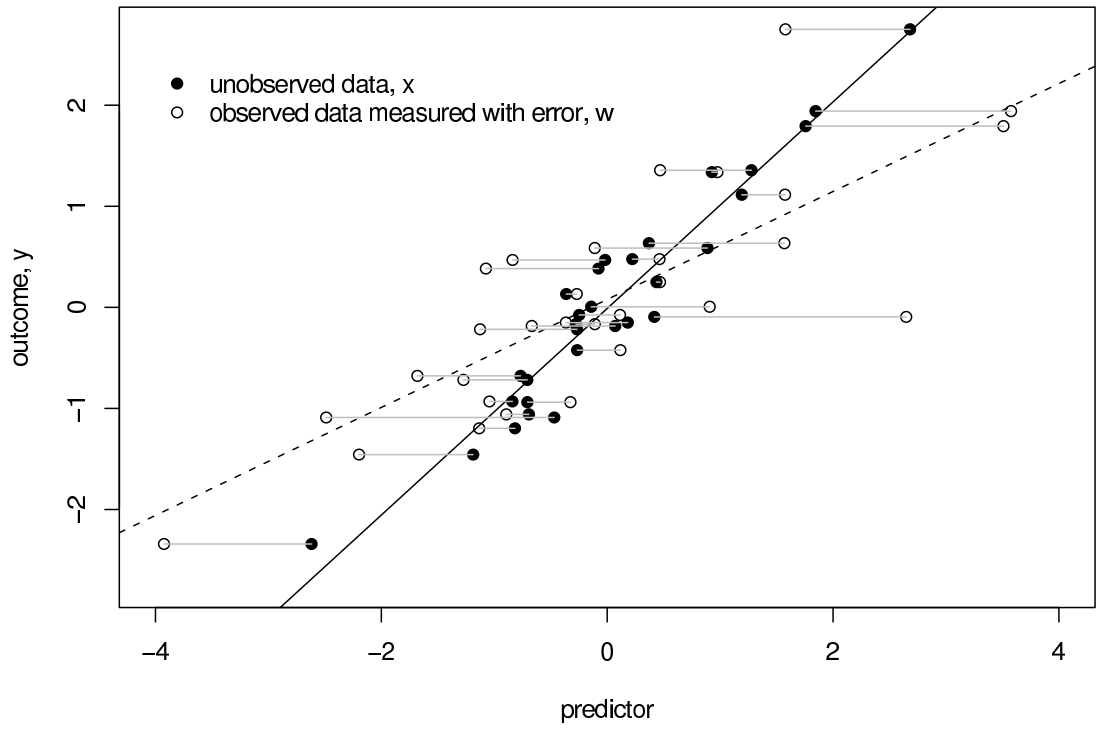


Figure 1.1: Attenuation due to classical measurement error in a simulated linear regression setting

1.2.2 Loss of power

We illustrate the loss of power using again a simple linear regression model, $y = \beta_0 + \beta_1 x + \epsilon$, where $x \sim N(\mu_x, \sigma_x^2), \epsilon \sim N(0, \sigma_\epsilon^2)$ and $u \sim N(0, \sigma_u^2)$. We assume that the measurement error u is independent of x and ϵ and the error variances are known. To test the hypothesis of no association, $H : \beta_1 = 0$ versus the alternative hypothesis, $H : \beta_1 > 0$ for a 1 percent α-level test, power equal to 90 percent, a true slope $\beta_1 = 1$ and error variances $\sigma_x^2 = \sigma_u^2 = \sigma_\epsilon^2 = 1$, the formula for the required sample size is

$$n = (Q_{N(0,1)}^{1-0.01} + Q_{N(0,1)}^{0.90})^2 \times \sigma_\epsilon^2/(\beta_1^2 \times \sigma_x^2),$$

$$\text{where } P(Z \leq Q_{N(0,1)}^{1-\alpha}) = \alpha \text{ and } Z \sim N(0,1).$$

The sample size required is then $n = (3.608)^2 \times 1/(1^2 \times 1) \approx 13$. When there is measurement error and reliability λ, the formula becomes

$$n = (Q_{N(0,1)}^{1-0.01} + Q_{N(0,1)}^{0.90})^2 \times (\sigma_\epsilon^2 + \lambda\beta_1\sigma_u^2)/(\lambda^2\beta_1^2 \times (\sigma_x^2 + \sigma_u^2))$$

and the sample size requirement for the same power and alpha-level inflates to $n = (3.608)^2 \times (1 + 0.5 \cdot 1 \cdot 1)/(0.5^2 \cdot 1^2 \times (1 + 1)) \approx 39$. The sample size triples. Figure 1.2 illustrates that when measurement error is present, it is necessary to acquire more sample information to achieve a prescribed level of power in the absence of measurement error.

1.3 Auxiliary data

To remedy the problem of bias, we require auxiliary data that typically comes in the form of validation or replicate data. These data can be used to estimate the variance of the measurement error. Replicate data experiments can, for example, capture measurement error variation in ultrasound, MRI, and CT measurements. Multiple readings can be taken over a single scan with different readers (addressing the error due to reader), multiple scans and single readers (addressing the error due

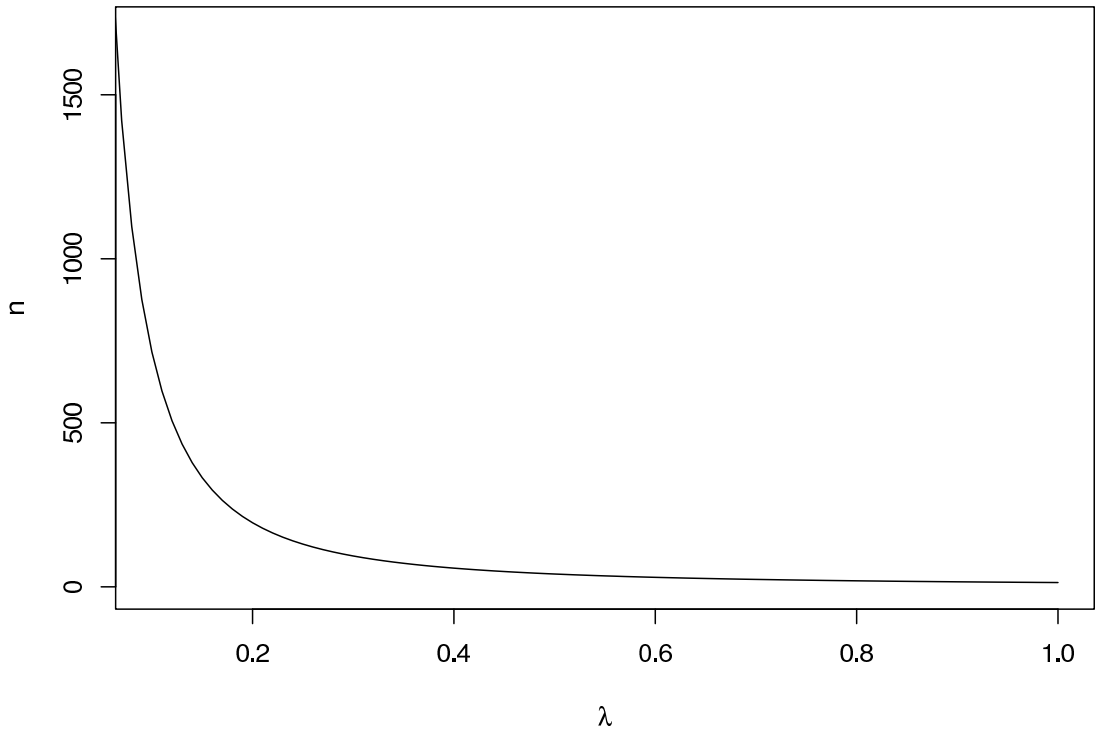


Figure 1.2: Sample size required to maintain 90% power for a given reliability and 1% α-level test, with slope and error variances of 1

to technology), or multiple scans and multiple readers (to address multiple sources of measurement error simultaneously). If replicate data are not available but prior knowledge of the reliability or estimates of measurement error are available, then other bias correction techniques may be applied. These methods are described in Chapter 2.

There are times when the measurements of interest can be obtained without error. These data are known as validation data. When these data are available on a subset of the sample, they can be used to correct for measurement error bias. However, there are times when obtaining validation data may not be feasible due to cost, or they may not be available. For example, if one is interested in obtaining measures of carotid IMT, the true measure can only be obtained by incision into the carotid artery.

Obtaining these "true" values could only be done ethically on autopsy. In contrast, in the genetics setting, one may view a genotyped SNP (following data cleaning) as a "true" value, but an imputed SNP as a value subject to error. This type of validation data is commonly available on a subset of study participants. The bias correction methods customized for this type of genetic validation data are presented in Chapter 3 and assessed in Chapter 4.

1.4 Extensions of standard correction techniques

We will also investigate extensions of traditional measurement error correction techniques to predictors with nonlinear error, such as that which occurs when considering a gene by gene interaction or gene by environment interaction. These extensions would also be suitable for nonlinear functional forms of a mismeasured predictor(s), such as, log-transformed C-reactive protein or ankle-brachial index (ABI), a ratio of two blood pressures. Applying standard measurement correction methods directly to these nonlinear functions of mismeasured predictors would violate underlying error model assumptions. A brief presentation is included in Chapter 3.

1.5 Looking ahead

In Chapter 2, we will review some of the popular correction methods from measurement error literature. In Chapter 3, we will present how these existing methods and extensions can be performed with error-prone imputed genotypes. In Chapter 4, we will apply these methods to genetic data from the Multi-Ethnic Study of Atherosclerosis (MESA) and assess their performance. Finally, in Chapter 5, we will present concluding remarks regarding this investigation.

Chapter 2

MODELS AND METHODS TO ADDRESS MEASUREMENT ERROR

In this chapter, we will present algebraic derivations of analytic results from measurement error literature. We start with presentation of the classical error model to address single and multiple mismeasured predictors in simple linear regression and multiple regression.

2.1 Measurement Error Models

Beginning with the classical error model, in this case true values are measured with additive error and usually a constant variance. Continuing with the notation introduced in Chapter 1, the observed value (W_{ij}) is the sum of the true value (X_i) and measurement error (U_{ij}), that is

$$\begin{aligned} W_{ij} &= X_i + U_{ij}, \text{ where} \\ X_i &\sim (\mu_x, \sigma_x^2), \\ U_{ij} &\sim (0, \sigma_u^2) \text{ with } E(U_{ij} \mid X_i) = 0, \text{ and} \\ X_i &\perp\!\!\!\perp U_{ij}, \text{ with } i = 1, \ldots, n \text{ and } j = 1, \ldots, k. \end{aligned}$$

Classical measurement error also has the property that the observed values vary more than the true values. Blood pressures, serum cholesterol, and carotid IMT, for example, tend to fall in this category.

Although, the classical model does not require that the X_i and U_{ij} follow normal distributions, this is a common assumption and we will follow this in our initial presentation of methods.

2.2 Simple Linear Regression

Suppose we are interested in investigating the linear association between Y and X by performing standard ordinary least squares (OLS) regression of the form

$$Y_i = \beta_0 + \beta_X X_i + \epsilon_i.$$

However, if W is observed and treated as X, as we saw in Chapter 1, the analyses would lead to biased estimates.

2.2.1 Method of Moments

In this approach $\hat{\beta}_W$ from OLS would not yield a consistent estimate of β_X, but instead a consistent estimate of $\lambda\beta_X$, where $\lambda = \frac{\sigma_x^2}{\sigma_w^2}$ is the attenuation factor, or reliability ratio. It is known that $\hat{\beta}_{X_{OLS}}$ is a consistent estimator of $\frac{\sigma_{y,x}}{\sigma_x^2}$. Therefore $\hat{\beta}_{W_{OLS}}$ is a consistent estimator of $\frac{\sigma_{y,w}}{\sigma_w^2}$, and it follows that:

$$\begin{aligned}
\hat{\beta}_{W_{OLS}} &= \frac{\hat{\sigma_{y,w}}}{\hat{\sigma_w^2}} \\
\text{and } \frac{\sigma_{y,w}}{\sigma_w^2} &= \frac{cov(Y,W)}{var(W)} \\
&= \frac{cov(Y,X+U)}{var(X+U)} \\
&= \frac{cov(Y,X)}{var(X+U)} \\
&= \frac{\sigma_{y,x}}{\sigma_w^2} \\
&= \frac{\sigma_x^2}{\sigma_w^2} \cdot \frac{\sigma_{y,x}}{\sigma_x^2} \\
\text{which estimates} \quad & \frac{\sigma_x^2}{\sigma_w^2}\beta_X \\
&= \lambda\beta_X.
\end{aligned}$$

This result motivates a method-of-moments type of bias correction in the case where an estimate of the measurement error variance is available, e.g.,

$$\hat{\beta}_{X_{MOM}} = \frac{\hat{\beta}_{W_{OLS}}}{\hat{\lambda}}$$

$$= \hat{\beta}_{W_{OLS}} \cdot \frac{\widehat{\sigma_x^2 + \sigma_u^2}}{\hat{\sigma_x^2}}$$
$$= \hat{\beta}_{W_{OLS}} \cdot \frac{\hat{\sigma_w^2}}{\hat{\sigma_w^2} - \hat{\sigma_u^2}}.$$

2.2.2 Regression Calibration

In using the method-of-moments correction, if the intercept β_0 is also of interest, bias correction would also be necessary on $\hat{\beta}_{0_{OLS}}$. So let's consider another perspective,

$$\begin{aligned} E(Y \mid W) &= E(E(Y \mid X, W) \mid W) \\ &= E(E(Y \mid X) \mid W) \\ &= E(E((\beta_0 + \beta_X X + \epsilon) \mid X) \mid W) \\ &= E((\beta_0 + \beta_X X) \mid W) \\ &= \beta_0 + \beta_X E(X \mid W). \end{aligned} \tag{2.1}$$

This result suggests that regressing on an estimate of $E(X \mid W)$ rather than W would lead to an unbiased estimate of β_X. We designate the corrected values of the predictor as the conditional mean of X given W.

If we assume X and U are normally distributed, we could derive the conditional mean from the multivariate normal distribution. However, the conditional mean holds more generally. It is the best linear predictor of Y that minimizes the mean square error (Carroll et al., 2006, pp. 361-363). Thus, we choose

$$E(X \mid W) = \mu_x + \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}(w - \mu_w)$$

In order to estimate the conditional mean $E(X \mid W)$, the measurement error variance, σ_u^2, must be estimated. Auxiliary data are required to estimate this variance component. For example, one may be able to obtain $\hat{\sigma_u^2}$ from literature on a particular predictor of interest, in which case corrected estimates can be obtained from the above

result. In the case of k replicates,

$$\begin{aligned} \hat{X}_i &= \hat{E}(X_i \mid \bar{W}_i = \bar{w}_i) \\ &= \hat{\mu}_x + \frac{\hat{\sigma_x^2}}{\hat{\sigma_x^2} + \hat{\sigma_u^2}/k}(\bar{w}_i - \hat{\mu}_w) \\ &= (1 - \lambda_k)\hat{\mu}_x + \lambda_k \bar{w}_i \end{aligned} \tag{2.2}$$

This estimate of X can be thought of as a weighted average of the data observed with error, w (or $\bar{w}$ in the case of replicates), and the mean of all observed values, $\hat{\mu}_x$. The weighted average tends towards values w when reliability λ is large and tends towards a constant vector of $\hat{\mu}_x$ when λ is small. In this situation, notice that greater measurement error ($\hat{\sigma_u^2}$) leads to greater attenuation by way of smaller λ.

This idea of regressing Y on the corrected values $\hat{X}$, is the motivation behind regression calibration methods. Finally, we will need to apply a correction to the standard error estimates of $\hat{\beta_X}$ by way of either bootstrap or large sample (sandwich) estimation methods.

We note that although it was shown in Chapter 1 that a decrease in power is observed for an uncorrected analysis, the Type I error rate is preserved. In this simple linear regression case this is because we have $H_0 : \beta_X = 0 \Leftrightarrow H_0 : \lambda\beta_X = 0$.

2.3 Multiple Linear Regression

2.3.1 Single mismeasured predictor

Now consider the case where we want to adjust for an additional covariate. We add to the previous model, covariate Z, which is measured precisely and where $Z_i \sim (\mu_z, \sigma_z^2)$. In this case, the model becomes

$$Y_i = \beta_0 + \beta_X X_i + \beta_Z Z_i + \epsilon_i.$$

Further, if X is correlated with Z, attenuation of the coefficient of W is more extreme and bias in the coefficient of Z is also observed. In this situation, the coefficient of

W estimates $\lambda^*\beta_X$, where

$$\lambda^* = \frac{\sigma^2_{x|z}}{\sigma^2_{w|z}} = \frac{\sigma^2_{x|z}}{\sigma^2_{x|z} + \sigma^2_u}.$$

The value of λ^* attains the value of λ if X and Z are uncorrelated, otherwise $\lambda^* < \lambda$ because $\sigma^2_{x|z} < \sigma^2_x$ (Carroll et al., 2006, p. 52).

2.3.2 Multiple mismeasured predictors

Now consider the case where we may have more than one mismeasured predictor and more than one precisely measured covariate. In this case, the model becomes

$$\mathbf{Y} = \beta_0 + \beta^t_X\mathbf{X} + \beta^t_Z\mathbf{Z} + \epsilon.$$

We define the following covariance matrices

$$\begin{aligned} \Sigma_{ww} &= Cov(W) \\ \Sigma_{uu} &= Cov(U) \\ \Sigma_{zz} &= Cov(Z) \\ \Sigma_{wz} = \Sigma^t_{zw} &= Cov(W, Z). \end{aligned}$$

The ordinary least squares regression coefficient estimates can be represented as

$$\begin{pmatrix} \hat{\beta}_{W_{OLS}} \\ \hat{\beta}_{Z_{OLS}} \end{pmatrix} = \begin{pmatrix} \Sigma_{ww} & \Sigma_{wz} \\ \Sigma_{zw} & \Sigma_{zz} \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_{ww} - \Sigma_{uu} & \Sigma_{wz} \\ \Sigma_{zw} & \Sigma_{zz} \end{pmatrix} \begin{pmatrix} \beta_X \\ \beta_Z \end{pmatrix}.$$

This bias evident in the OLS estimators could be corrected as

$$\begin{pmatrix} \hat{\beta}_X \\ \hat{\beta}_Z \end{pmatrix} = \begin{pmatrix} \Sigma_{ww} - \Sigma_{uu} & \Sigma_{wz} \\ \Sigma_{zw} & \Sigma_{zz} \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_{ww} & \Sigma_{wz} \\ \Sigma_{zw} & \Sigma_{zz} \end{pmatrix} \begin{pmatrix} \hat{\beta}_{W_{OLS}} \\ \hat{\beta}_{Z_{OLS}} \end{pmatrix}.$$

Further, to apply regression calibration, Carroll and Stefanski (1990) derived the following results

$$\hat{E}(\mathbf{X}_i \mid \mathbf{Z}_i, \bar{\mathbf{W}}_{i\cdot}) = \hat{\mu}_w + (\hat{\Sigma}_{xx}, \hat{\Sigma}_{xz}) \begin{pmatrix} \hat{\Sigma}_{xx} - \hat{\Sigma}_{uu}/k & \hat{\Sigma}_{xz} \\ \hat{\Sigma}_{zx} & \hat{\Sigma}_{zz} \end{pmatrix}^{-1} \begin{pmatrix} \bar{\mathbf{W}}_{i\cdot} - \hat{\mu}_w \\ \mathbf{Z}_i - \bar{\mathbf{Z}}_{\cdot} \end{pmatrix}. \quad (2.3)$$

Note that again, if there are no replicates available, or $k = 1$, we would need to substitute an external estimate for Σ_{uu} to proceed. When k has a value of 1 or 2, regression calibration produces identical estimates to the method-of-moments estimates (Carroll et al., 2006, pp. 71-72). This result also holds when there is a nonconstant number of replicates, denoted by substituting k_i for k.

2.4 Alternate Error Structures

Other biomedical predictors of interest may be subject to a Berkson additive error model. In the Berkson error model, true values (X_i) are composed of the observed values (W_i) and an error term (U_i) added to it, $X_i = W_i + U_i$.

One distinction between classical and Berkson errors is in the variability of the observed values relative to the variability of the true values. In contrast to classical error models, in Berkson error models, the observed values have less variation than the true values (Carroll et al., 2006, pp. 5, 27-28). Examples such as dosiometry calculations for radiation exposure are generally assumed to fall under this structure.

Further, measurement error models can be even more complicated than either of these simple models. For instance, in some cases the observed measures are not unbiased measures of the true values. In these situations, the classical model does not hold, but Kipnis et al. (1999, 2001, 2003) have described a class of models to allow for such bias.

Chapter 3

REGRESSION CALIBRATION EXTENSIONS FOR GENOTYPE IMPUTATION DATA

In this chapter, our primary focus is customization of regression calibration for genetic validation data. This still relies on estimating the quantity $\hat{X} = \hat{E}(X \mid W)$. For this problem we have validation auxiliary data and can consider a more general error model. We begin this chapter with an introduction to genotype imputation, follow with a detailed presentation of the technique for single SNP models, and end with an extension for gene by gene interaction models.

3.1 Genotype imputation

The process of imputation and estimating genotypes introduces a component of error. Imputation makes use of underlying reference panels, most commonly the International HapMap (Thorisson et al., 2005; Frazer et al., 2007) and 1,000 Genomes Projects (Nielsen, 2010; Altshuler et al., 2010). These are international research collaborations that have resulted in creation of public databases of human genetic variation among worldwide populations.

There are a variety of genotype imputation programs available. Some of the more popular programs include IMPUTE, MACH, and BEAGLE, each with a varying underlying algorithm (Marchini et al., 2007; Howie et al., 2009; Li et al., 2009, 2010; Browning and Browning, 2007, 2009; Hao et al., 2009; Huang et al., 2012). For example, IMPUTE version 2 first requires phasing of the study genotypes, the process of separating the alleles according to parental sequences. Each parental sequence is considered a haplotype. The phasing is achieved by using an extended hidden Markov

model (HMM) for each individual's genotypes and assigning probability weights to haplotypes in cases of phase uncertainty through iterated steps of Markov chain Monte Carlo (MCMC). The algorithm alternates between phasing and haploid imputation conditional on the estimated haplotypes. The unique algorithm uses local similarities to obtain a customized reference panel from which to impute each haplotype of each study individual (Howie et al., 2009; Marchini and Howie, 2008, 2010).

A genotype for a particular SNP can be represented in various ways. Typical datasets from genotyping centers provide genotypes formatted in terms of DNA bases, for example, "CC", "CT", or "TT", for a specific SNP and individual. However, in conducting statistical analyses, it may be more convenient to recode this to a dosage format in terms of a specified or coded allele. For a genotyped SNP, the dosage would take on a value of 0, 1, or 2. If "C" is our coded allele, the three above genotypes would have dosages of 2, 1, and 0, respectively. Imputation software such as IMPUTE produces posterior probabilities of 3 possible genotypes, say AA, AB, and BB, where A and B are "dummy" alleles. In this case, the dosage corresponding to an additive model is computed as $\frac{0 \cdot p_{AA} + 1 \cdot p_{AB} + 2 \cdot p_{BB}}{p_{AA} + p_{AB} + p_{BB}}$ and interpreted as the estimated number of copies of the coded "dummy" allele B. Notice the imputed value is no longer confined to discrete values, though is still bounded by 0 and 2.

To get a sense of the degree of extrapolation that comes with imputation, the most recent consortium recommendations are for participating groups to impute to 40 million variants. Of these 40 million variants, typically no more than one million are genotyped. The initial focus is on p-values from each test of association to determine a small number of "top hits" and report effect sizes among those few SNPs. Particularly, to account for possible detection of false signals among these millions of tests, SNPs with p-value $< 5 \cdot 10^{-8}$ are considered as genome-wide significant, while SNPs with p-value $< 5 \cdot 10^{-5}$ often warrant further investigations. This raises the question of whether the error associated with SNP imputation warrants use of a measurement error correction in these types of analyses, as we know potential biases and losses in

power may present. Further, following the initial models with single SNPs, subsequent interest may be in models with multiple SNPs, such as, a gene by gene interaction or gene by environment interaction. In either case, the gene and environment variables may be subject to measurement error.

Little research has been done to study imputation error. One group compares approaches for accounting for uncertainty introduced by imputation, but doesn't consider modeling the error structure. The authors compare OLS regression on the "best-guess" imputed genotype (discrete values), regression on the "dosage" (semicontinuous), and "mixture regression models" that account for the individual posterior probabilities from imputation (Zheng et al., 2011). The authors find that all models yield similar power for non-modest minor allele frequencies (MAF) when there is a true additive effect. Further, power is found to be dependent on imputation accuracy. When there is a true dominant effect, dependency of power on allele frequency is even more pronounced. However, when considering practicality and accuracy, the authors find method 2, using the dosage is suitable in "most realistic settings".

3.2 Implementing regression calibration for single SNP models

Here, we take a different approach to account for the uncertainty introduced by imputation. Specifically, we propose that genotype imputation introduces a component of error and hypothesize that such error can lead to biased regression coefficients and loss of power in tests. As a result, we consider applying regression calibration to imputed SNPs.

For this proposed measurement error correction, it is necessary to have the SNP of interest genotyped on a subset of participants (i.e., validation data). As shown in Chapter 2, the correction is motivated by the replacement of X by an estimate of $E(X \mid W)$, or in the case of covariates, $E(X \mid W, Z)$. After estimating the conditional mean, one can proceed with a standard analysis.

3.2.1 Notation and model of interest

First, we review the underlying model and test of interest. We let X be the dosage of a given genotyped SNP which is assumed as perfectly measured after data cleaning. To test for an association between the SNP and phenotype, we use a model (below) that adjusts for covariates Z_k, age, gender, field center of recruitment, and includes only those from a common racial or ethnic group (e.g. Caucasians). The model is

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \Sigma_k \beta_k Z_{ki} + \epsilon_i, \\ &\quad \text{for } i = 1, \ldots, n \text{ and } k = 2, \ldots, K, \end{aligned} \tag{3.1}$$

where, the ith person has phenotype y_i, genotype x_i, age z_{2i}, gender z_{3i}, and field center category indicators z_{4i}, z_{5i}, $\ldots$, z_{Ki}. The β_0, β_1, and the β_k are unknown parameters. Primary interest is in testing $H_0 : \beta_1 = 0$, which corresponds to no association between SNP and phenotype, accounting for differences in age, gender, and field center. The parameter β_1 represents the difference in mean levels of the phenotype when comparing groups of individuals with one copy versus two copies (or zero versus one copy) of the coded allele for a particular SNP, given individuals are the same age and gender, and have the same field center of recruitment.

However, in practice, the majority of SNPs in a GWAS are not genotyped. Most SNPs are imputed but analyzed as though they were precisely measured. Here, W is the dosage of an imputed SNP which is observed with imputation error.

$$\begin{aligned} Y_i &= \beta_0^* + \beta_1^* W_i + \Sigma_k \beta_k^* Z_{ki} + \epsilon_i^*, \\ &\quad \text{for } i = 1, \ldots, n \text{ and } k = 2, \ldots, K. \end{aligned}$$

After fitting the OLS regression corresponding to the naive model, the primary test of interest instead tests $H_0 : \beta_1^* = 0$. Due to the error present in the imputed genotypes

W_i, the naive model parameters may also differ, and hence are designated as β_0^*, β_1^*, and the β_k^*. We observed in Chapter 2 that although this test is still an α level test, since $H_0 : \beta_1^* = 0 \Leftrightarrow H_0 : \lambda^* \beta_1 = 0 \Leftrightarrow H_0 : \beta_1 = 0$, a decrease in power and bias in the SNP coefficient estimate are still expected. As a result, we propose the following regression calibration correction.

3.2.2 The underlying error model

For many imputed SNPs, auxiliary data is available which may take the form of replication data or validation data. Many studies continually repeat the imputation process on new reference panels. Currently, there is disagreement whether additional iterations yield "better" imputation, and instead the iterations are considered as yielding alternative datasets, e.g., replication data. In other situations, genotyping may have occurred on multiple genotyping chips or arrays and on only a subset of individuals. This is common for Candidate Gene studies which preceeded the GWAS era. Candidate Gene studies are a potential source of validation data on a subset of individuals. We focus our proposed correction on validation data.

In the validation setting, X is observed only on a subset S of individuals and is unobserved on the rest. Recall from section 2.2.2 that regression calibration is driven by a step where X is replaced by an estimate of $E(X \mid W)$, or in the case of covariates $E(X \mid W, Z)$. We designate the corrected values of the predictor as $\hat{X} = \hat{E}(X \mid W, Z)$.

Since X and W are correlated and W is known, X can be predicted from W, with best linear predictor of the form

$$X_i = \gamma_0 + \gamma_1 W_i + \Sigma_k \gamma_k Z_{ki} + U_i. \tag{3.2}$$

This is the typical form of a regression calibration model as it focuses on X given (W, Z).

3.2.3 Regression calibration of imputed SNPs

With respect to the regression calibration model in equation 3.2, in the subset S, we regress X on the error-prone predictor W and other covariates Z to obtain estimates $\hat{\gamma_0}$, $\hat{\gamma_1}$, and the $\hat{\gamma_k}$. The coefficient estimates can then be used to obtain predicted values based on the imputed values and covariates. These predicted values can be considered as the calibrated imputed values, $\hat{x}$, for individuals without genotyping on the SNP of interest. For the subset of individuals with available genotyping, the values x can be retained. (Carroll et al., 1990, 2006, pp. 65-66, 70; Thurston et al., 2003.)

$$\hat{x_i} = \begin{cases} x_i, & \text{if } i \in S \\ \hat{\gamma_0} + \hat{\gamma_1} w_i + \Sigma_k \hat{\gamma_k} z_{ki}, & \text{if } i \notin S \end{cases}$$

Finally, we fit the original model of interest with $\hat{X}$ in place of X and perform the desired test.

$$Y_i = \beta_0^{**} + \beta_1^{**} \hat{X}_i + \Sigma_k \beta_k^{**} Z_{ki} + \epsilon_i^{**}$$

As with replication-based correction, validation-based regression calibration requires accounting for the additional estimation of unknown genotypes by correcting standard errors of estimated coefficients via bootstrap or sandwich estimators (Huber, 1967; White, 1982). Section C.1 of Appendix 3 includes annotated R code which implements each of the steps of validation-based regression calibration. The output displays results from uncorrected and corrected approaches as well as the gold standard.

3.3 Implementing regression calibration for SNP interaction models

Now suppose we want to apply regression calibration to a model with a gene by gene interaction. The first SNP will be denoted with subscript a, and the second SNP

denoted with subscript b. As before, X represents the genotyped values, measured precisely, while W denotes the imputed values measured with error.

The method presented here is in parallel with that presented for the single SNP model. Each SNP is calibrated individually, then the desired regression is performed substituting the calibrated values $(\hat{X}_a, \hat{X}_b)$ for (X_a, X_b).

3.3.1 Notation and model of interest

We consider the following linear model is of interest and specifically the test of $H_0 : \beta_3 = 0$, whether the coefficient of the interaction term is different from zero.

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{ai} + \beta_2 X_{bi} + \beta_3 X_{ai} X_{bi} + \Sigma_k \beta_k Z_{ki} + \epsilon_i \\ &\quad \text{for } i = 1, \ldots, n \text{ and } k = 2, \ldots, K. \end{aligned} \tag{3.3}$$

As before, by assuming the imputed values are precisely measured and carrying out the standard analysis, this naive model is instead considered:

$$\begin{aligned} Y_i &= \beta_0^* + \beta_1^* W_{ai} + \beta_2^* W_{bi} + \beta_3^* W_{ai} W_{bi} + \Sigma_k \beta_k^* Z_{ki} + \epsilon_i^* \\ &\quad \text{for } i = 1, \ldots, n \text{ and } k = 2, \ldots, K. \end{aligned} \tag{3.4}$$

3.3.2 The underlying error model

In order to estimate $\hat{X}_a$ and $\hat{X}_b$, we consider the regression calibration models

$$X_{ai} = \gamma_0^{'} + \gamma_1^{'} W_{ai} + \gamma_2^{'} W_{bi} + \gamma_3^{'} W_{ai} W_{bi} + \Sigma_k \gamma_k^{'} Z_{ki} + U_i^{'}, \tag{3.5}$$

$$X_{bi} = \gamma_0^{''} + \gamma_1^{''} W_{ai} + \gamma_2^{''} W_{bi} + \gamma_3^{''} W_{ai} W_{bi} + \Sigma_k \gamma_k^{''} Z_{ki} + U_i^{''}, \tag{3.6}$$

$$\text{where } E(U^{'} \mid W, Z) = 0 \text{ and } E(U^{''} \mid W, Z) = 0.$$

3.3.3 Regression calibration with SNP interaction

Similar to the single SNP case, we regress each of X_a and X_b on all covariates (**W**,**Z**) among the subset of individuals with genotyping. After fitting the regression calibration model, we can calibrate the imprecisely measured values as follows:

$$\hat{x}_{ai} = \begin{cases} x_{1i}, & \text{if } i \in S \\ \hat{\gamma_0}' + \hat{\gamma_1}' w_{ai} + \hat{\gamma_2}' w_{bi} + \hat{\gamma_3}' w_{ai} w_{bi} + \Sigma_k \hat{\gamma_k}' z_{ki}, & \text{if } i \notin S \end{cases} \tag{3.7}$$

$$\hat{x}_{bi} = \begin{cases} x_{2i}, & \text{if } i \in S \\ \hat{\gamma_0}'' + \hat{\gamma_1}'' w_{ai} + \hat{\gamma_2}'' w_{bi} + \hat{\gamma_3}'' w_{ai} w_{bi} + \Sigma_k \hat{\gamma_k}'' z_{ki}, & \text{if } i \notin S \end{cases} \tag{3.8}$$

The corrected values that result are the predicted values, except where genotypes are available (in subset S). For those cases, genotypes are used in place of the fitted values.

Finally, we can fit the OLS regression using the calibrated values.

$$Y_i = \beta_0^{**} + \beta_1^{**} \hat{X}_{ai} + \beta_2^{**} \hat{X}_{bi} + \beta_3^{**} \hat{X}_{ai} \hat{X}_{bi} + \Sigma_k \beta_k^{**} Z_{ki} + \epsilon_i^{**}$$

As with the single SNP models, we will then need to account for the additional estimation of unknown genotypes by correcting standard errors of estimated coefficients via bootstrap or sandwich estimators. Annotated R code is included in Section C.2 of Appendix 3 which details our comparison of validation-based regression calibration with an interaction of two imputed SNPs (using sandwich standard error estimates).

Chapter 4

APPLICATION TO THE MULTI-ETHNIC STUDY OF ATHEROSCLEROSIS

In this application, we aim to assess whether measurement error correction through extensions of regression calibration can improve on the standard methods using an uncorrected genotype dosage, to best account for imputation uncertainty when considering additive genetic models. We consider data from the Multi-Ethnic Study of Atherosclerosis (MESA) SNP Health Association Resource (SHARe) (Bild et al., 2002; CHSCC 2012) which aims to identify genetic variants underlying subclinical cardiovascular disease and progression, as well as risk factors that predict disease progression. Subclinical disease is characterized by non-invasive detection of disease before clinical signs are apparent.

As part of the SHARe effort (NHLBI 2010, 2011), 934,148 SNPs on 8298 participants were genotyped on the Affymetrix 6.0 chip. After standard genotyping quality control filtering, data remained for 854,755 SNPs on 8227 participants. Imputation has been completed as two rounds for SHARe, the first using HapMap 1+2 (3.9 million SNPs) and the second using the 1000 Genomes pilot dataset (12-17 million SNPs) (Howie et al., 2011), each considered as an alternative dataset for the other. The number of SNPs includes those genotyped in SHARe and is before post-imputation QC and SNP filtering. A third round of imputation is underway using the 1000 Genomes integrated dataset, which will produce datasets with 39,295,245 variants. As part of the CARe effort, 48,982 SNPs were genotyped on the IBC chip on 6482 MESA participants. A more complete summary of MESA imputation and genotyping efforts is given in Table 4.1.

Using this available genetic data for MESA participants, we are able to apply our proposed measurement error corrections. SHARe individuals contribute imperfectly measured imputed SNPs while CARe individuals contribute precisely measured genotyped SNPs, yielding validation data. In fact, separate corrections can be applied to each round of imputation. Alternatively, a combined correction can be applied to the two rounds if paired and thought of as replicates, and irrespective of the CARe genotypes. However, replication-based correction requires additional exploration of underlying assumptions and will be the focus of future work. This chapter will assess the use of validation-based measurement error correction in this GWAS setting.

For this example, we consider serum triglycerides as the phenotype of interest. Levels of triglycerides, or blood fats, have a role in evaluation and management of CVD risk (Miller et al., 2011). This particular phenotype is relatively well-studied, with several implicated SNP associations from literature (Aulchenko et al., 2009; Chasman et al., 2008; Kathiresan et al., 2008, 2009; Kooner et al., 2008; Lamina et al., 2011; Sabatti et al., 2009; Saxena et al., 2007; Tan et al., 2012). We focus our attention on 24 genotyped SNPs in MESA CARe, of which 23 are imputed in MESA SHARe.

Following the investigation of single SNP associations, we consider gene by gene interactions which may account for unattributed genetic determinants of lipid levels. Despite the success in identifying SNPs associated with CAD and risk factors such as triglycerides, much of the genetic component of CAD remains unattributed (Lanktree and Hegele, 2009). Accounting for gene by gene (g $\times$ g) interactions may contribute to the missing heritability in complex disease (Cordell, 2009). A few studies have reported plausible interaction effects (Tam et al., 2009; Tan et al., 2012). Here we focus on one study (the Bogalusa Heart Study) which has reported an interaction effect of two SNPs (genotyped and imputed in MESA) on serum triglyceride levels in young adults (Xin et al., 2003).

In both single SNP and g $\times$ g interaction models, we also exclude individuals

taking lipid-lowering medications. In the genetic setting, we usually model each race/ethnic group separately. This leaves us with 2026 Caucasians on which we focus our analyses. A small number of participants were removed such that the 2026 participants have complete data for the 23 SNPS, both CARe genotypes and SHARe imputed values. Restricting to this group allows us to better assess how well the proposed methods perform compared to analysis of the true data. We consider blinding experiments where we vary the size of the subset (or equivalently, the percentage) with validation data. That is, for this subset the true genotypes are known for all 2026 participants (for the 23 SNPs of interest), but for the sake of experimentation we pretend it is known for only p% of the 2026 participants. We obtain the calibration equation (as written in equation 3.2) based on the subset with validation data and apply the equation to the imputed values.

If the corrections prove to be promising, they could be useful in current work in MESA as well as other groups. While CARe IBC provides a source of validation data for 93% of the Caucasians and 99% of the Chinese, it covers roughly only two-thirds in the African-American and Hispanic groups as seen in Table 4.1. Additionally, Candidate Gene Rounds 1+2 and the current Metabochip data present another opportunity to apply the proposed methods with even further reduced validation coverage on largely unique sets of genotyped SNPs. This is especially appealing because the Candidate Gene data was slightly overshadowed by the availability of SHARe data due to the timing of advances in genotyping technology. As a result, investigators made minimal use of the Candidate Gene data on roughly 700 participants and a small number of SNPs (3,070) with the expanse of MESA SHARe data just around the corner. Hence, the proposed methods would allow the 700 participants with genotyping in candidate regions to inform the association analysis between imputed genotypes and phenotype. The proposed correction would potentially yield more power to detect subtle associations than using either the Candidate Gene or SHARe dataset alone.

Table 4.1: MESA imputation and genotyping summary

	CAU	CHN	AFA	HIS	total	number of variants	
Available imputation: SHARe (includes genotyping)						CAU	non-CAU
HM1+2 gen1	2685	777	2588	2174	8224	3,891,070	
1000G gen2	2685	777	2588	2174	8224	12,143,325	17,406,968
1000G gen3	2685	777	2588	2174	8224	39,295,245	39,295,245

	CAU	CHN	AFA	HIS	total	number of variants
Currently available genotyping						
SHARe	2687	777	2658	2176	8298	934,148
CARe IBC	2500	771	1751	1460	6482	48,982
CG 1+2	712	718	712	705	2487	3,070
CG 3	2487	754	1689	1448	6378	766
Metabochip	0	0	580	0	580	196,725

CAU = Caucasian

CHN = Chinese

AFA =African-American

HIS = Hispanic

In this setting, X represents any SNP genotyped on the CARe IBC chip and passing QC. W represents the dosage of an imputed SNP which is observed with imputation error. The uncorrected values correspond to W and the corrected values correspond to $\hat{X}$. We will consider both examples, imputation using HapMap 1+2 (HM1+2) as well as 1000 Genomes (1000G). In this chapter, in order to assess performance, we present results when using the gold standard (genotyped dosage), uncorrected imputed dosage (HM1+2 and 1000G), and corrected dosage (HM1+2 and 1000G, after measurement error correction using the techniques previously described).

4.1 Nature of the SNP dosage data

Figures A.1 and A.2 (Appendix A) show scatter plots of the genotyped CARe IBC SNP versus the imputed HapMap 1+2 SNP, both expressed as the dosage of the coded allele. Each point represents one participant. The dashed black line is the reference line, $x = w$. The solid red line results from a regression, fitting x versus w. The quality score (ratio of observed to expected variances of the imputed dosage), $oevar$, is also provided.

Most of the regression lines correspond well by eye with the reference line in Figures A.1 and A.2. Only a few SNPs have regression lines that deviate moderately from the reference line, despite some poor quality ratios. Specifically, three SNPs in Figure A.1 have $oevar < 0.50$ and three additional SNPs in Figure A.2 have $oevar < 0.55$. It is a somewhat surprising to notice the magnitude of spread of imputed values relative to each level of genotype, e.g., homozygous minor, homozygous major, and especially heterozygous. Such large spread is evident even for imputed SNPs with higher quality ratios, for example, rs7694035 with $oevar = 0.77$ in Figure A.2.

Similarly, Figures A.3 and A.4 show scatter plots of the genotyped CARe IBC SNP versus the imputed 1000 Genomes SNP, again, both expressed as the dosage of the coded allele. Notice the negative slope for some of the scatter plots with imputed genotypes using the 1000 Genomes panel. This is almost certainly the result

of differences in SNP strand annotation. For the purposes of our analyses, we will assume these are coding issues resulting from strand flips which are not of interest for the purposes of assessing measurement error. As a result, we will invert the dose of the 1000 Genomoes SNP (inverted dosage = 2 - original dosage), to correspond to the noninverted CARe IBS dose, whenever the slope of the regression of x versus w is negative, or equivalently when the correlation is negative. Although 1000 Genomes strand information may be more current, it is more convenient to invert for only one panel. Notice, in these cases, the reference lines correspond to $x = 2 - w$ in Figures A.3 and A.4.

Figures A.5 and A.6 show scatter plots comparing the two sources of imputation. The correlation is noted with each SNP as well as imputation quality ratios. Again, where the correlation is negative, the reference line and 1000G dosages are inverted and assumed due to a difference in strand annotation. As with the previous figures, here the dashed black line is the reference line, $w_2 = w_1$ or $w_2 = 2 - w_1$, and the solid red line is the regression line from fitting w_2 versus w_1, (1000G versus HapMap 1+2 imputed SNP).

4.2 Results of validation-based regression calibration in single-SNP models

We present results using the known data from the CARe genotyping (x), as well as uncorrected estimates from both sources of imputation (w_1 and w_2), compared with validation-based regression calibration ($\hat{x_1}$ and $\hat{x_2}$) for each of the 23 SNPs. Regression output is presented where 40, 50, or 90% of the participants have known genotypes and the remaining 60, 50, or 10% use the calibrated dosage in place of the imputed dosage (Table B.1). In Appendix B, the subscript s denotes results based on sandwich standard error estimates, while the subscript m denotes model-based standard error estimates. Sandwich-based results were very similar to bootstrap-based results (not shown). We consider a level of significance of 0.05 because each of these 23 SNPs is

previously implicated as having an association with triglyceride level (Aulchenko et al., 2009; Chasman et al., 2008; Kathiresan et al., 2008, 2009; Kooner et al., 2008; Lamina et al., 2011; Sabatti et al., 2009; Saxena et al., 2007; Tan et al., 2012). If instead this were a discovery phase, a level of $5 \cdot 10^{-8}$ is commonly used.

A number of SNPs illustrate the potential value in the validation-based correction, even when only 40 or 50% of the data have known values. Specifically, note rs439401 on the second page of Table B.1, in which the reduction of bias and change in standard errors is evident after applying the proposed measurement error correction. This is particularly true of the 1000 Genomes imputation, where having 40 or 50% known genotypes gives a result much closer to that of the known genotypes compared to the uncorrected imputed data (Figure 4.1).

Table 4.2. presents a summary of the results from Table B.1 (Appendix B). Tabulation of the total number of SNPs with significant associations detected (out of 23) is provided for the analyses of known genotypes as well as uncorrected and corrected values. Multiple entries of the number of significant hits in corrected analyses are provided to get an idea of the dependence of the results on each random subset with known genotypes (and to a lesser extent due to the bootstrap standard errors). Row $\hat{x}_p$ corresponds to a blinding experiment where only p percent have known values. The mean and median absolute bias is also assessed for the uncorrected and corrected analyses. Absolute bias is summarized over one of these experiments but across all 23 SNPs.

In this dataset with 2026 participants, the corrected values $\hat{x}_p$ perform very well relative to the standard analysis of the uncorrected values (in terms of significant hits). It appears that having 40-50% of individuals in the study genotyped yields corrected analyses that can detect comparably as many hits as having the full cohort genotyped. In terms of both mean and median absolute bias across the 23 SNPs, there is a clear improvement over the uncorrected analysis when as little 30% of the genotypes are known. Specifically, when the 2026 CARe genotypes are known, 15 of

the 23 SNPs reach statistical significance. In comparison, when only HM1+2 imputed (uncorrected) genotypes are used, only 6 SNPs reach statistical significance when model-based standard errors are used and 11 when sandwich standard errors are used. However, in the measurement error corrected approach, when 40% of participants have known genotypes, our handful of random subsets yield 13 to 15 significant hits, nearly as many as the gold standard (with genotypes on all participants known). Not surprisingly, mean and median absolute bias show a decreasing trend as the number of individuals with known genotypes increases (p increases).

It is of interest to examine the two-way and three-way relationships among absolute bias, the percent p with known genotypes, and imputation quality score *oevar*. Figure 4.2 shows bias versus p on the SNP-level and each of HM1+2 and 1000G imputation. As expected, the range of bias decreases with p among the corrected analyses. Note, $p = 0$ represents uncorrected analyses on imputed values. Figures 4.3 and 4.4 provide a better view of the distribution of bias for each value of p and separately for the HM1+2 and 1000G imputation. We observe that the bias values cluster more tightly around zero, as p increases. In Figure 4.5, we examine the previous relationship by *oevar*. This is of interest because studies will often decide to filter out all results on SNPs with *oevar* below a certain threshold. It appears there may be some bias improvement in the coefficient estimates for SNPs with *oevar* between 0.7 and 0.8 when p=0.4 or 0.5. SNPs with *oevar* in this range may be filtered out and not used in analysis due to poor imputation quality.

In summary, these preliminary results could be of interest in future study design in terms of a potential cost reduction. Current GWAS may be improved in terms of ability to detect signals due to application of the correction. In our dataset, the measurement error corrected approach (with as little as 40% genotyped) yields comparably as many detected signals as when the full cohort is genotyped. Further, this could lead to fuller utilization of imputed data with only moderate quality which would ordinarily be removed from analysis.

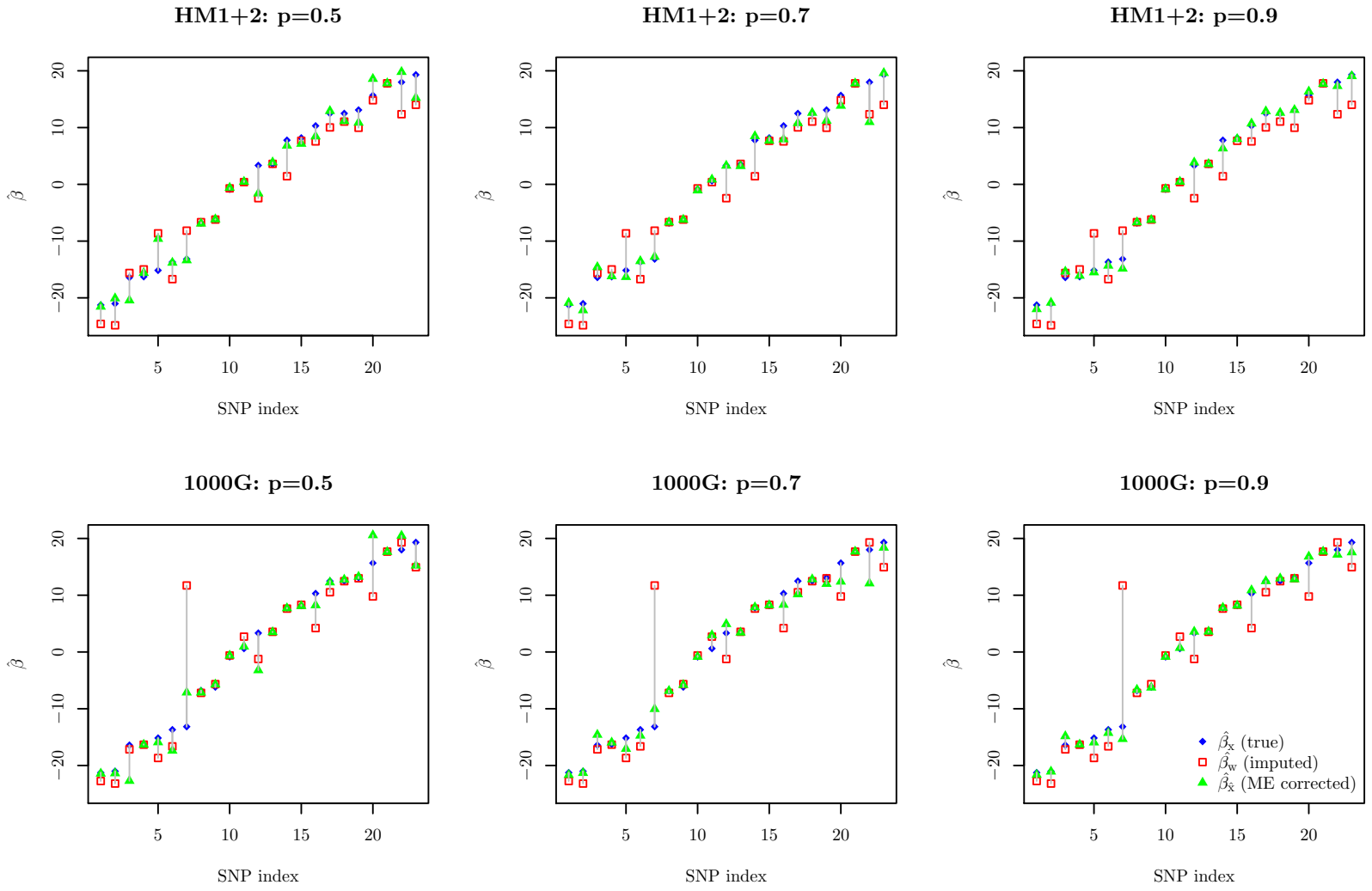


Figure 4.1: Comparison of true coefficient estimates with those from uncorrected and corrected analysis

Table 4.2: Comparison of corrected, uncorrected, and gold standard analysis. Number of significant associations ($p < 0.05$) and absolute bias from single SNP analyses. Subscripts m, s, and b denote model-based, sandwich, and bootstrap standard errors, respectively.

predictor	no. of significant hits		absolute bias (mg/dL)			
	p-value < 0.05		HM1+2		1000G	
	HM1+2	1000G	mean	median	mean	median
Using true values (CARe genotypes), the known "gold standard":						
x_m	15	15	0	0	0	0
x_s	15	15	0	0	0	0
Using imputed dosage with no correction:						
w_m	6	8	2.554	2.463	2.770	2.770
w_s	11	11	2.554	2.463	2.770	2.770
w_b	10	10	2.554	2.463	2.770	2.770
Measurement error corrected based on p% genotyped individuals:						
$\hat{x}_{b\cdot 10}$	11, 10, 11, 11	11, 11, 14, 10	2.405	1.213	3.119	1.181
$\hat{x}_{b\cdot 20}$	13, 12, 13, 12	12, 12, 13, 11	2.674	1.061	1.833	1.142
$\hat{x}_{b\cdot 30}$	12, 13, 13, 11	12, 14, 14, 13	1.968	1.384	1.660	0.483
$\hat{x}_{b\cdot 40}$	15, 13, 14, 15	14, 14, 15, 15	1.883	1.174	1.866	0.470
$\hat{x}_{b\cdot 50}$	14, 13, 14, 13	15, 14, 14, 15	1.496	0.937	1.726	0.355
$\hat{x}_{b\cdot 60}$	15, 14, 15, 13	16, 16, 16, 14	1.626	1.163	1.586	0.849
$\hat{x}_{b\cdot 70}$	16, 16, 15, 17	15, 16, 16, 17	0.987	0.388	1.278	0.925
$\hat{x}_{b\cdot 80}$	12, 14, 15, 15	14, 15, 15, 16	0.758	0.412	0.689	0.341
$\hat{x}_{b\cdot 90}$	16, 15, 15, 15	16, 15, 16, 16	0.430	0.240	0.493	0.249

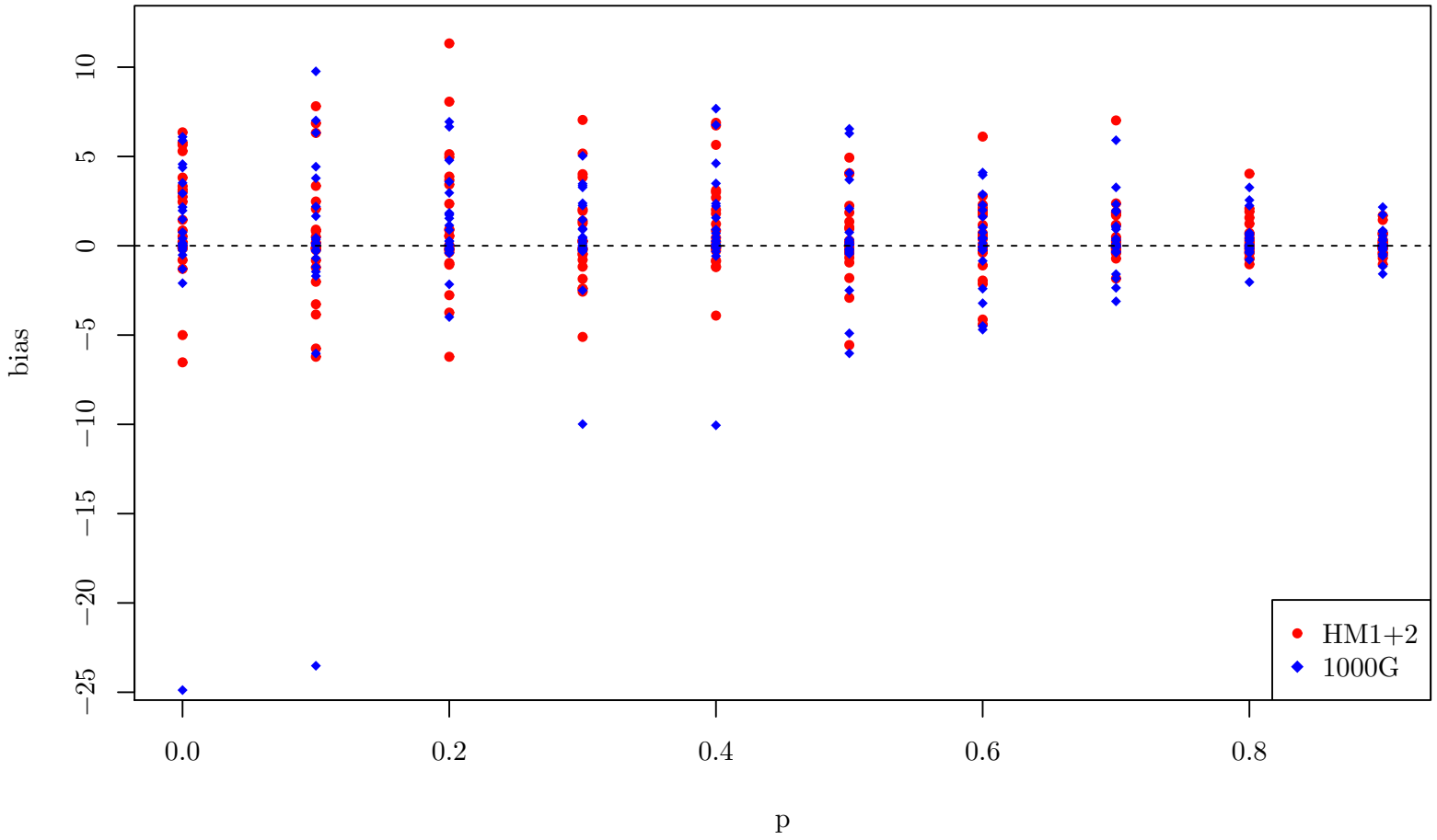


Figure 4.2: Bias versus p (all oevar)

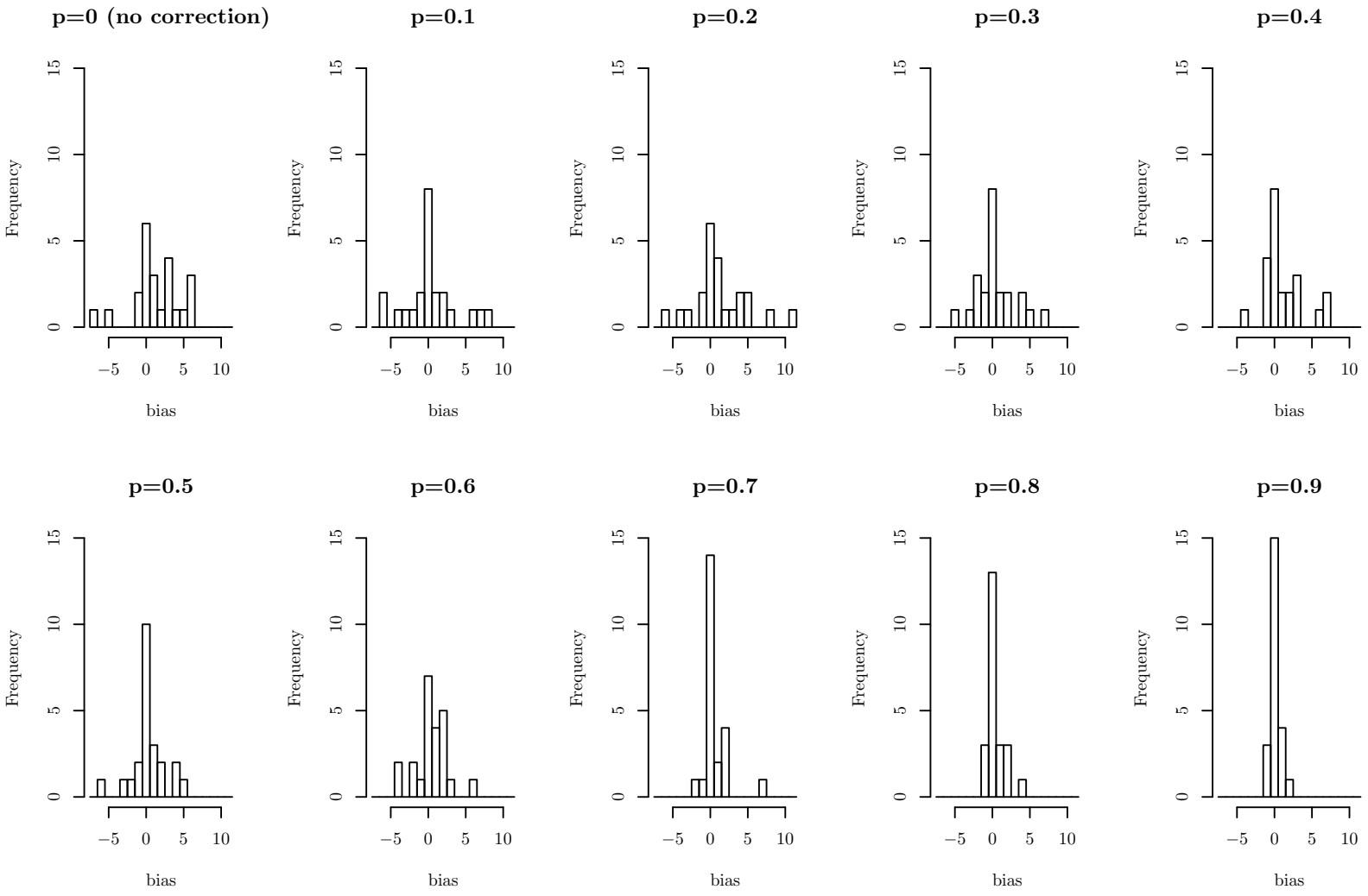


Figure 4.3: HapMap 1+2: Histogram of bias by p

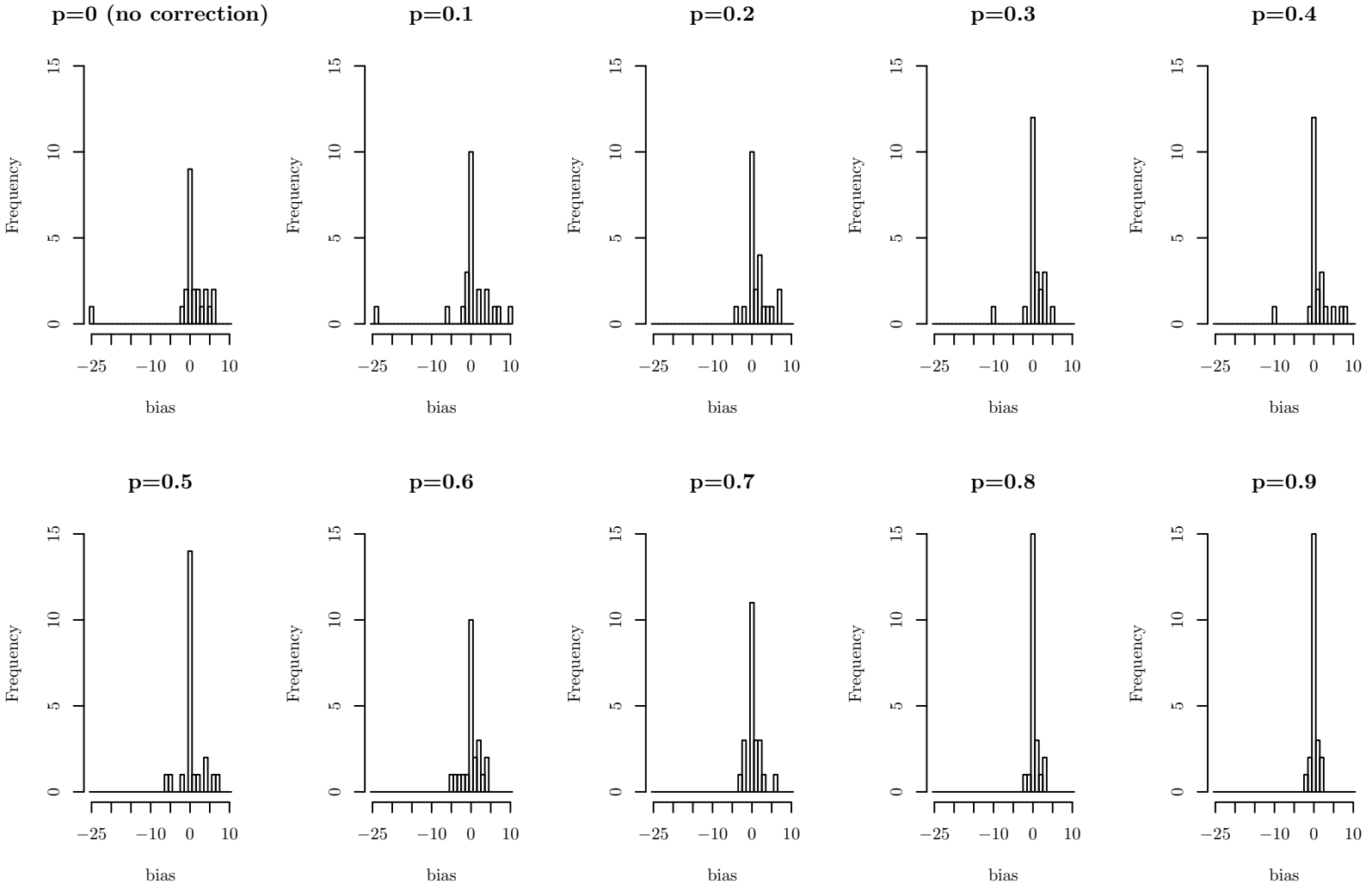


Figure 4.4: 1000 Genomes: Histogram of bias by p

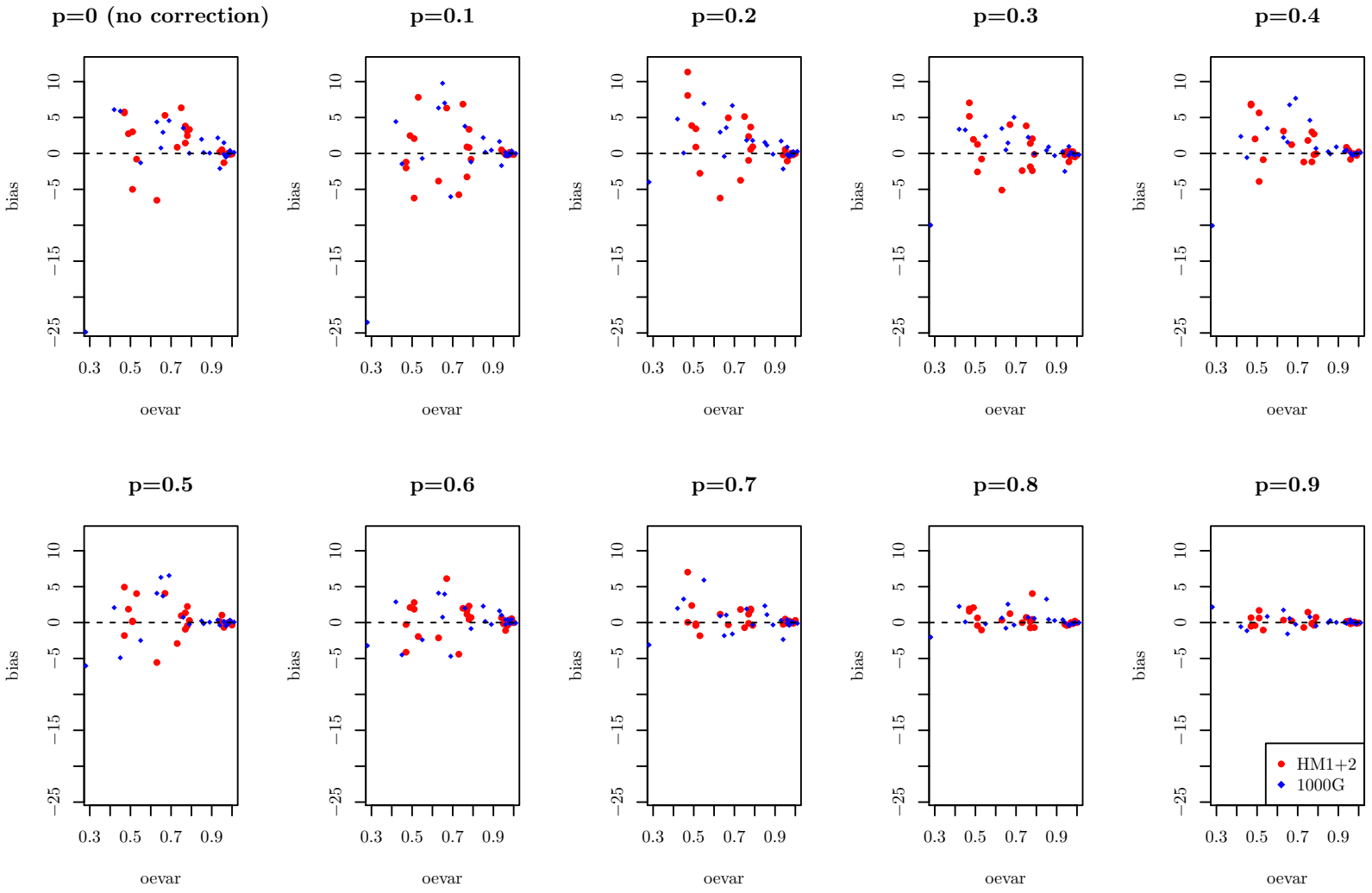


Figure 4.5: Bias versus oevar (by p)

4.3 Consideration of standard errors of SNP coefficient estimates

4.3.1 Impact of type of standard error on results

Skewness and heteroscedasticity are present in the triglycerides data. That is, the triglycerides values have differing variances for different values of genotype dosage. Consequently, the method used to compute standard errors of estimated coefficients impacts the number of detected signals among the uncorrected analyses of w's. If instead, the natural logarithm of the triglycerides value is taken, model-based standard errors are very comparable to both sandwich and bootstrap standard errors. As a result, the number of detected signals is similar when using any of the three types of standard errors with the transformed phenotype. Here, our results are presented in terms of the untransformed values for ease of biological interpretation.

Nonetheless, it is important to note the differences when model-based standard errors are used. Genetic analysis software without means of obtaining either sandwich or bootstrap standard errors is quite common in GWAS. Genetic analysis software typically implements special handling of memory which dramatically speeds up analyses. However, the specialized software often does not allow for great flexibility or incorporation of options such as sandwich standard errors. Use of non-model-based standard errors might improve the ability to detect associations, as we have seen, but often is not done due to the difficulty in handling and analyzing of such large files.

4.3.2 Bias versus variance

Recall from Chapter 2 that in the case of simple linear regression, the method of moments estimator of β_{SNP} is equivalent to that from the regression calibration approach. Hence, we have:

$$\hat{\beta}_{\hat{X}} = \frac{\hat{\beta}_W}{\hat{\lambda}}$$

$$\begin{aligned} V(\hat{\beta}_{\hat{X}}) &= \frac{1}{\hat{\lambda}^2} \cdot V(\hat{\beta}_W) \\ &= \frac{[\hat{V}(W)]^2}{[\hat{V}(X)]^2} \cdot V(\hat{\beta}_W) \end{aligned}$$

With imputation data, it may or may not be the case that $V(W) > V(X)$, which is expected under classical measurement error where it is assumed $W = X + U$ and $V(W) = V(X) + V(U)$. With imputed dosage, W and genotyped dosage, X, we often observe instead that $\hat{V}(W) < \hat{V}(X)$, due to a typically negative $\hat{Cov}(X, U)$. As a result, the standard error estimates of $\hat{\beta_{\hat{X}}}$ may be smaller than that of $\hat{\beta_W}$ as is often seen in Tables 4.3, 4.4, and B.1 (in the Appendix).

This observation is in contrast to the usual bias versus variance tradeoff characteristic of corrected estimators. That is, that the nature of bias correction leads to an estimator which is more variable than the original biased estimator, the cost for reduced bias (Carroll et al., 2006, pp. 60-63, 69). Our atypical observation is not surprising due to the discrete nature of the genotyped values and continuous nature of the imputed values over the same range.

4.4 Results of validation-based regression calibration in SNP-interaction models

The first section of Table 4.3 shows the estimates corresponding to the main effect of rs328 (g_a). Line 1 is obtained using known genotypes (x_a); lines 2-4 are obtained using HapMap 1+2 imputed genotypes ($w_{1a}, \hat{x}_{1a \cdot p}$); and lines 5-7 are obtained using the 1000 Genomes imputation ($w_{2a}, \hat{x}_{2a \cdot p}$). The second and third sections show the estimates corresponding to the main effect of rs1800588 (g_b) and the interaction term (g_{ab}) in the same 7 models as section 1.

As in the single SNP analyses, we vary the size of the random subset, where row $\hat{x}_p$ reflects the results corresponding to a blinding experiment where only p percent have known values. Results are presented in Table 4.3 for values of $p = 0.5$ and 0.9.

Table 4.3: Assessing a g×g interaction of g_a = rs328 and g_b = rs1800588
Bootstrap SEs used for 4 corrected analyses ($\hat{x}_p$) and model-based SEs used for gold standard (x) and 2 uncorrected analyses (w_1, w_2).

	Estimate	Std. Error	95% CI or TI		Pr(>| t |)
x_a	17.024	5.792	5.671	28.376	0.003
w_{1a}	17.952	5.754	6.674	29.229	0.002
$\hat{x}_{1a\cdot 50}$	17.360	4.612	8.776	26.847	$1.7 \cdot 10^{-4}$
$\hat{x}_{1a\cdot 90}$	17.379	4.402	9.070	26.693	$7.9 \cdot 10^{-5}$
w_{2a}	18.446	5.754	7.169	29.723	0.001
$\hat{x}_{2a\cdot 50}$	17.726	4.625	8.926	27.344	$1.3 \cdot 10^{-4}$
$\hat{x}_{2a\cdot 90}$	17.324	4.405	8.960	26.679	$8.4 \cdot 10^{-5}$
x_b	6.156	14.792	−22.837	35.148	0.677
w_{1b}	8.044	14.868	−21.097	37.186	0.589
$\hat{x}_{1b\cdot 50}$	6.488	10.111	−12.591	27.353	0.521
$\hat{x}_{1b\cdot 90}$	6.862	9.828	−10.718	28.275	0.485
w_{2b}	11.231	14.748	−17.676	40.138	0.446
$\hat{x}_{2b\cdot 50}$	8.314	10.013	−11.476	29.358	0.406
$\hat{x}_{2b\cdot 90}$	6.607	9.802	−10.895	27.699	0.500
x_{ab}	1.145	8.078	−14.689	16.978	0.887
w_{1ab}	−0.138	8.158	−16.127	15.851	0.987
$\hat{x}_{1ab\cdot 50}$	0.593	6.123	−12.567	12.177	0.923
$\hat{x}_{1ab\cdot 90}$	0.794	5.924	−11.312	11.503	0.893
w_{2ab}	−1.592	8.063	−17.395	14.211	0.843
$\hat{x}_{2ab\cdot 50}$	−0.242	6.097	−13.138	11.378	0.968
$\hat{x}_{2ab\cdot 90}$	1.006	5.918	−11.160	11.656	0.865

Although both SNPs are significant in the single SNP models, only the main effect of rs328 is significant in the model containing both main effects and interaction term, even when using the known genotypes from CARe. Particularly, with all three terms in the model adjusted for basic demographics, the standard errors are very large relative to the coefficient estimates for the main effect of rs1800588 and the interaction term.

Although we are unable to replicate the finding of the Bogalusa Heart Study (Xin et al., 2003), the impact of the measurement error correction can be clearly seen in this example. For the main effect for rs328, both sets of uncorrected imputed data do fairly well, though less bias is observed when the measurement error corrected imputed values are used. For the main effect of rs1800588, coefficient estimates based on the uncorrected imputed data are quite biased away from the null, especially for the 1000G imputation (effect size of 6.2 for true data compared with 11.2 for 1000G). However, the coefficient estimates are much less biased using the values obtained by applying the proposed correction (effect size of 6.5 and 6.9 for HM1+2 and 8.3 and 6.6 for 1000G, when 50% and 90% of validation data is available).

In Table 4.3, all corrected analyses use bootstrap standard errors, while uncorrected analyses use model-based standard errors. For comparison, Table 4.4 repeats the same 7 analyses but instead using sandwich standard errors across all analyses. We notice that within each term for this example, the standard errors are much more uniform. However, the differences in bias are still observed. Also note the coefficient estimates using the values with the proposed correction applied ($\hat{x}_p$) depend on the random subsample chosen. This is why the coefficient estimates differ in the corrected analyses across Tables 4.4 and 4.5, even though using different standard error estimators should have no impact.

The analysis was repeated in the CARe data with the natural log-transformed triglycerides value, with and without potentially sparse categories recoded. The results were very similar to the untransformed analysis, with p-values for the interaction

term larger than 0.85.

Unfortunately, we are unable to replicate the findings of the Bogalusa study here, but there are some major differences between the populations considered. The Bogalusa Heart Study was comprised of 65% non-Hispanic Caucasians and 35% non-Hispanic African-Americans from Bogalusa, LA. The study consisted of 6 cross-sectional exams every 3 years between 1978 and 1996. The first of the exams included young adults of age 18 to 20, while the final exam included individuals of age 18-41. In the genetic analysis of interest, the authors considered a subset including only those with blood samples collected for DNA genotyping (between 1988 and 1996). Of those, 1,291 individuals (70% Caucasian) had genotypes for rs328 and rs1800588 with a mean age of 25.8. Roughly 31% had one exam, 27% had two exams, and 42% had three to seven exams.

The authors used mixed models to analyze the multiple observations per person, considering the natural log-transformed triglycerides values and adjusting for age, age^2, gender, race, BMI, and BMI^2 (centered). The main finding was based on the combined sample including both Caucasians and African-Americans (though genotype and allele frequency differences between the races were assessed).

They reported that the interaction term was significant in the combined sample of 1,291, but only borderline among the Caucasians only (2307 observations on 902 individuals), and not-significant among the African-Americans only (918 observations on 389 individuals).

It was also noted that the triglycerides measures (following a 12 hour fast) came from two different procedures, before versus after 1986 but meet acceptable performance criteria in the CDC-LSP (Centers for Disease Control and Prevention - Lipid Standardization Program, 2012).

In contrast, the analysis performed on the subset of 2,026 MESA SHARe participants were aged 44 to 84 (mean 62.1) and free of clinical CVD at the baseline exam from which the triglycerides values were measured, between 2000 and 2002. The par-

Table 4.4: Assessing a g×g interaction of g_a = rs328 and g_b = rs1800588 Sandwich SEs used for all 7 models.

	Estimate	Std. Error	95% CI or TI		Pr(>| t |)
x_a	17.024	4.431	8.339	25.709	$1.2 \cdot 10^{-4}$
w_{1a}	17.952	4.488	9.156	26.748	$6.3 \cdot 10^{-5}$
$\hat{x}_{1a \cdot 50}$	17.429	4.481	8.646	26.213	$1.0 \cdot 10^{-4}$
$\hat{x}_{1a \cdot 90}$	17.361	4.420	8.697	26.024	$8.6 \cdot 10^{-5}$
w_{2a}	18.446	4.481	9.664	27.229	$3.8 \cdot 10^{-5}$
$\hat{x}_{2a \cdot 50}$	17.873	4.455	9.141	26.605	$6.0 \cdot 10^{-5}$
$\hat{x}_{2a \cdot 90}$	17.428	4.426	8.754	26.103	$8.2 \cdot 10^{-5}$
x_b	6.156	9.963	−13.371	25.683	0.537
w_{1b}	8.044	10.145	−11.840	27.928	0.428
$\hat{x}_{1b \cdot 50}$	7.779	10.076	−11.970	27.528	0.440
$\hat{x}_{1b \cdot 90}$	7.856	10.054	−11.849	27.561	0.435
w_{2b}	11.231	10.004	−8.376	30.838	0.262
$\hat{x}_{2b \cdot 50}$	9.359	9.949	−10.142	28.860	0.347
$\hat{x}_{2b \cdot 90}$	7.966	9.928	−11.493	27.425	0.422
x_{ab}	1.145	6.036	−10.686	12.975	0.850
w_{1ab}	−0.138	6.156	−12.204	11.928	0.982
$\hat{x}_{1ab \cdot 50}$	0.397	6.098	−11.556	12.349	0.948
$\hat{x}_{1ab \cdot 90}$	0.207	6.067	−11.685	12.098	0.973
w_{2ab}	−1.592	6.083	−13.516	10.331	0.794
$\hat{x}_{2ab \cdot 50}$	−0.393	6.067	−12.283	11.498	0.948
$\hat{x}_{2ab \cdot 90}$	0.029	6.012	−11.754	11.811	0.996

ticipants were recruited from six regions in the United States: Baltimore City and Baltimore County, Maryland; Chicago, Illinois; Forsyth County, North Carolina; Los Angeles County, California; New York, New York; and St. Paul, Minnesota (Bild et al., 2002).

Hence, the differences in the characteristics between the Bogalusa and MESA participants as well as differences in analysis methods may explain why we could not replicate the results in MESA. Further exploration is desired in assessing improvement in using our proposed correction method when there is interest in gene by gene interactions between imputed SNPs.

Chapter 5

DISCUSSION AND CONCLUSIONS

5.1 General observations from uncorrected analyses

In the previous chapter we began our investigation examining the concordance between true genotypes and imputed genotypes in each of 23 SNPs which are thought to be associated with triglyceride levels. The degree of noise seen in our plots was somewhat surprising. We also looked at the concordance between imputed genotypes when using different reference panels (namely, HapMap 1+2 and 1000 Genomes pilot) in each of the implicated SNPs which was also quite noisy. Further, comparing inference when using true genotypes on the full sample versus uncorrected imputed values, we observed a reduction in the number of detected signals (previously implicated) as well as introduction of bias in the coefficient estimates of interest.

5.2 Advantages of using regression calibration on imputed genotypes

Previous work by other groups recommended use of an imputed dosage over a best guess genotype in order to account for imputation uncertainty. It is generally viewed that by using the dosage, measurement error due to imputation is adequately addressed. However, our investigation shows that further improvements can be made when regression calibration is applied to imputed genotypes rather than performing OLS regression without correction. Such improvements include reduction of bias in coefficient estimates and increased power to detect signals. Further, due to the nature of the error, we typically do not experience a large cost in increased variance of the corrected estimator, if any, associated with the reduction in bias.

Given validation data, one may be tempted to conduct standard analysis within

the subset of (size p*2026) individuals with known genotypes and forgo a corrected approach. However, when using just these participants, we do not observe the same degree of concordance among the signals as when using the proposed correction on the imputed data. A quick examination of this approach yields results with greater bias in coefficient estimates as well as a larger reduction in power than when the proposed correction is used. This finding (of course) depends on both the size of the full sample as well as size of the subset. A brief examination included subsets of up to 70% with known genotypes of the 2026 total participants.

We also assessed transportability of the calibration equation that is derived from 50% of known genotypes and taking corrected values as strictly the fitted values. Although the scenario may be artificial, the purpose is to assess whether the regression calibration equation has value for studies with similar populations, or possibly in other races. It is noted that the correction would be expected to perform better on this dataset, the data from which the calibration equation came from rather than new data. The impact here would be on magnitude of bias in the coefficient estimates rather than number of significant hits. The calibration equation applied to all participants only offers one column as the linear combination of other columns of the design matrix, which does not change p-values.

In summary, we believe this correction has valuable practical use. As highlighted in the previous chapter, there are many examples of validation genotype data in MESA with a range of validation coverage (percentage of known values). These sets of validation data can be used to inform tests of association between imputed genotype and phenotype yielding results that are closer to those based on the full set of true genotypes (usually unknown) than when analyzing either the known subset or imputed values alone. Similar secondary genotyping on subsets of participants is not uncommon in other studies.

5.3 Limitations of the methodology

The current methods require validation data, that is genotypes on a subset of the participants in addition to the imputed values. Not surprisingly, the degree of success attained by the correction is correlated with the size of the subset. However, it is promising that improvement is seen even when the relative size of the subset is quite small.

Our initial findings are based on one particular phenotype which might yield effect sizes which are relatively larger than other complex traits of interest. In addition, the results from each corrected model is dependent on the random subset with known genotypes chosen (e.g., the blinding at random).

It is worth noting that our findings may be specific to an easier to detect effect in this well-studied and widely-available phenotype relative to other traits of interest. We should also consider the minor allele frequency (MAF), the underlying genetic model, and more carefully examine the effects of influential observations for triglycerides, as well as other phenotypes of interest. This dataset presents a unique opportunity to perform additional simulations to assess performance of the proposed correction methods under more general situations.

We had mentioned but not analyzed a 24th SNP in the previous chapter, which was genotyped in both CARe and SHARe. Cross tabulation of this SNP reveals that there is still some genotyping error in our "gold standard" which is ignored. For this SNP, we observed that 2007 of the 2026 individuals (99%) had concordant genotype calls across the two platforms, though 3 of the remaining 19 individuals with discordant calls had both alleles in disagreement (e.g. GG in CARe versus CC in SHARe). However, error due to imputation remains the more impactful component.

5.4 Future work

In parallel to our validation-based correction, we will next consider a replication-based correction where imputation based on different reference panels can be viewed as replicate data. We can still apply regression calibration methods motivated by substituting an estimate of $E(X \mid W, Z)$ for X. However, this will require some modification of the standard methods presented in Chapter 2 in order to address the departure from the classical error model. In this case, X is not observed, not even in a subset. Further, we may take a structural modeling approach by imposing distributional assumptions on X. This will be the focus of our upcoming work.

As an extension, we will also consider the absence of both validation and replicate data. In this case, measurement error can still be addressed by obtaining an estimate of the measurement error variances from external sources.

BIBLIOGRAPHY

D. Altshuler, R. M. Durbin, G. R. Abecasis, et al., and 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, Oct 2010.

Y. S. Aulchenko, S. Ripatti, I. Lindqvist, D. Boomsma, I. M. Heid, P. P. Pramstaller, B. W. Penninx, A. C. Janssens, J. F. Wilson, T. Spector, N. G. Martin, N. L. Pedersen, K. O. Kyvik, J. Kaprio, A. Hofman, N. B. Freimer, M. R. Jarvelin, U. Gyllensten, H. Campbell, I. Rudan, A. Johansson, F. Marroni, C. Hayward, V. Vitart, I. Jonasson, C. Pattaro, A. Wright, N. Hastie, I. Pichler, A. A. Hicks, M. Falchi, G. Willemsen, J. J. Hottenga, E. J. de Geus, G. W. Montgomery, J. Whitfield, P. Magnusson, J. Saharinen, M. Perola, K. Silander, A. Isaacs, E. J. Sijbrands, A. G. Uitterlinden, J. C. Witteman, B. A. Oostra, P. Elliott, A. Ruokonen, C. Sabatti, C. Gieger, T. Meitinger, F. Kronenberg, A. Doring, H. E. Wichmann, J. H. Smit, M. I. McCarthy, C. M. van Duijn, L. Peltonen, Y. S. Aulchenko, S. Ripatti, I. Lindqvist, D. Boomsma, I. M. Heid, P. P. Pramstaller, B. W. Penninx, A. C. Janssens, J. F. Wilson, T. Spector, N. G. Martin, N. L. Pedersen, K. O. Kyvik, J. Kaprio, A. Hofman, N. B. Freimer, M. R. Jarvelin, U. Gyllensten, H. Campbell, I. Rudan, A. Johansson, F. Marroni, C. Hayward, V. Vitart, I. Jonasson, C. Pattaro, A. Wright, N. Hastie, I. Pichler, A. A. Hicks, M. Falchi, G. Willemsen, J. J. Hottenga, E. J. de Geus, G. W. Montgomery, J. Whitfield, P. Magnusson, J. Saharinen, M. Perola, K. Silander, A. Isaacs, E. J. Sijbrands, A. G. Uitterlinden, J. C. Witteman, B. A. Oostra, P. Elliott, A. Ruokonen, C. Sabatti, C. Gieger, T. Meitinger, F. Kronenberg, A. Doring, H. E. Wichmann, J. H. Smit, M. I. McCarthy, C. M. van Duijn, and L. Peltonen. Loci influencing lipid levels and coronary

heart disease risk in 16 European population cohorts. *Nat. Genet.*, 41(1):47–55, Jan 2009.

D. E. Bild, D. A. Bluemke, G. L. Burke, R. Detrano, A. V. Diez Roux, A. R. Folsom, P. Greenland, D. R. Jacob, R. Kronmal, K. Liu, J. C. Nelson, D. O'Leary, M. F. Saad, S. Shea, M. Szklo, and R. P. Tracy. Multi-ethnic study of atherosclerosis: objectives and design. *Am. J. Epidemiol.*, 156(9):871–881, Nov 2002.

B. L. Browning and S. R. Browning. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.*, 84(2):210–223, Feb 2009.

S. R. Browning and B. L. Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.*, 81(5):1084–1097, Nov 2007.

R. J. Carroll and L. A. Stefanski. Approximate quasilikelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association*, 85:652 – 663, 1990.

R.J. Carroll, D. Ruppert, L.A. Stefanski, and C.M. Crainiceanu. *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*. Monographs on Statistics and Applied Probability. Taylor & Francis, 2006.

Centers for Disease Control and Prevention. Laboratory Quality Assurance and Standardization Programs: Lipid Standardization Program, August 2012. URL `http://www.cdc.gov/labstandards/lsp.html`.

D. I. Chasman, G. Pare, R. Y. Zee, A. N. Parker, N. R. Cook, J. E. Buring, D. J. Kwiatkowski, L. M. Rose, J. D. Smith, P. T. Williams, M. J. Rieder, J. I. Rotter, D. A. Nickerson, R. M. Krauss, J. P. Miletich, and P. M. Ridker. Genetic loci

associated with plasma concentration of low-density lipoprotein cholesterol, high-density lipoprotein cholesterol, triglycerides, apolipoprotein A1, and Apolipoprotein B among 6382 white women in genome-wide analysis with replication. *Circ Cardiovasc Genet*, 1(1):21–30, Oct 2008.

Collaborative Health Studies Coordinating Center, University of Washington, Seattle, WA. MESA SHARe - Study Overview, 2012. URL `http://www.mesa-nhlbi.org/MesaInternal/MESASHARe/MESASHARe.aspx`.

H. J. Cordell. Detecting gene-gene interactions that underlie human diseases. *Nat. Rev. Genet.*, 10(6):392–404, Jun 2009.

K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W. Belmont, A. Boudreau, P. Hardenbol, S. M. Leal, S. Pasternak, D. A. Wheeler, T. D. Willis, F. Yu, H. Yang, C. Zeng, Y. Gao, H. Hu, W. Hu, C. Li, W. Lin, S. Liu, H. Pan, X. Tang, J. Wang, W. Wang, J. Yu, B. Zhang, Q. Zhang, H. Zhao, H. Zhao, J. Zhou, S. B. Gabriel, R. Barry, B. Blumenstiel, A. Camargo, M. Defelice, M. Faggart, M. Goyette, S. Gupta, J. Moore, H. Nguyen, R. C. Onofrio, M. Parkin, J. Roy, E. Stahl, E. Winchester, L. Ziaugra, D. Altshuler, Y. Shen, Z. Yao, W. Huang, X. Chu, Y. He, L. Jin, Y. Liu, Y. Shen, W. Sun, H. Wang, Y. Wang, Y. Wang, X. Xiong, L. Xu, M. M. Waye, S. K. Tsui, H. Xue, J. T. Wong, L. M. Galver, J. B. Fan, K. Gunderson, S. S. Murray, A. R. Oliphant, M. S. Chee, A. Montpetit, F. Chagnon, V. Ferretti, M. Leboeuf, J. F. Olivier, M. S. Phillips, S. Roumy, C. Sallee, A. Verner, T. J. Hudson, P. Y. Kwok, D. Cai, D. C. Koboldt, R. D. Miller, L. Pawlikowska, P. Taillon-Miller, M. Xiao, L. C. Tsui, W. Mak, Y. Q. Song, P. K. Tam, Y. Nakamura, T. Kawaguchi, T. Kitamoto, T. Morizono, A. Nagashima, Y. Ohnishi, A. Sekine, T. Tanaka, T. Tsunoda, P. Deloukas, C. P. Bird, M. Delgado, E. T. Dermitzakis, R. Gwilliam, S. Hunt, J. Morrison, D. Powell, B. E. Stranger, P. Whittaker, D. R. Bentley, M. J. Daly, P. I. de Bakker, J. Barrett, Y. R. Chretien, J. Maller, S. McCarroll, N. Patterson, I. Pe'er, A. Price, S. Purcell,

D. J. Richter, P. Sabeti, R. Saxena, S. F. Schaffner, P. C. Sham, P. Varilly, D. Altshuler, L. D. Stein, L. Krishnan, A. V. Smith, M. K. Tello-Ruiz, G. A. Thorisson, A. Chakravarti, P. E. Chen, D. J. Cutler, C. S. Kashuk, S. Lin, G. R. Abecasis, W. Guan, Y. Li, H. M. Munro, Z. S. Qin, D. J. Thomas, G. McVean, A. Auton, L. Bottolo, N. Cardin, S. Eyheramendy, C. Freeman, J. Marchini, S. Myers, C. Spencer, M. Stephens, P. Donnelly, L. R. Cardon, G. Clarke, D. M. Evans, A. P. Morris, B. S. Weir, T. Tsunoda, J. C. Mullikin, S. T. Sherry, M. Feolo, A. Skol, H. Zhang, C. Zeng, H. Zhao, I. Matsuda, Y. Fukushima, D. R. Macer, E. Suda, C. N. Rotimi, C. A. Adebamowo, I. Ajayi, T. Aniagwu, P. A. Marshall, C. Nkwodimmah, C. D. Royal, M. F. Leppert, M. Dixon, A. Peiffer, R. Qiu, A. Kent, K. Kato, N. Niikawa, I. F. Adewole, B. M. Knoppers, M. W. Foster, E. W. Clayton, J. Watkin, R. A. Gibbs, J. W. Belmont, D. Muzny, L. Nazareth, E. Sodergren, G. M. Weinstock, D. A. Wheeler, I. Yakub, S. B. Gabriel, R. C. Onofrio, D. J. Richter, L. Ziaugra, B. W. Birren, M. J. Daly, D. Altshuler, R. K. Wilson, L. L. Fulton, J. Rogers, J. Burton, N. P. Carter, C. M. Clee, M. Griffiths, M. C. Jones, K. McLay, R. W. Plumb, M. T. Ross, S. K. Sims, D. L. Willey, Z. Chen, H. Han, L. Kang, M. Godbout, J. C. Wallenburg, P. L'Archeveque, G. Bellemare, K. Saeki, H. Wang, D. An, H. Fu, Q. Li, Z. Wang, R. Wang, A. L. Holden, L. D. Brooks, J. E. McEwen, M. S. Guyer, V. O. Wang, J. L. Peterson, M. Shi, J. Spiegel, L. M. Sung, L. F. Zacharia, F. S. Collins, K. Kennedy, R. Jamieson, and J. Stewart. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–861, Oct 2007.

K. Hao, E. Chudin, J. McElwee, and E. E. Schadt. Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies. *BMC Genet.*, 10:27, 2009.

B. Howie, J. Marchini, and M. Stephens. Genotype imputation with thousands of genomes. *G3 (Bethesda)*, 1(6):457–470, Nov 2011.

B. N. Howie, P. Donnelly, and J. Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.*, 5(6):e1000529, Jun 2009.

J. Huang, D. Ellinghaus, A. Franke, B. Howie, and Y. Li. 1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 Data. *Eur. J. Hum. Genet.*, 20(7):801–805, Jul 2012.

P.J. Huber. The behavior of maximum likelihood estimates under non-standard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley*, pages 221–233, 1967.

S. Kathiresan, O. Melander, C. Guiducci, A. Surti, N. P. Burtt, M. J. Rieder, G. M. Cooper, C. Roos, B. F. Voight, A. S. Havulinna, B. Wahlstrand, T. Hedner, D. Corella, E. S. Tai, J. M. Ordovas, G. Berglund, E. Vartiainen, P. Jousilahti, B. Hedblad, M. R. Taskinen, C. Newton-Cheh, V. Salomaa, L. Peltonen, L. Groop, D. M. Altshuler, and M. Orho-Melander. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat. Genet.*, 40(2):189–197, Feb 2008.

S. Kathiresan, C. J. Willer, G. M. Peloso, S. Demissie, K. Musunuru, E. E. Schadt, L. Kaplan, D. Bennett, Y. Li, T. Tanaka, B. F. Voight, L. L. Bonnycastle, A. U. Jackson, G. Crawford, A. Surti, C. Guiducci, N. P. Burtt, S. Parish, R. Clarke, D. Zelenika, K. A. Kubalanza, M. A. Morken, L. J. Scott, H. M. Stringham, P. Galan, A. J. Swift, J. Kuusisto, R. N. Bergman, J. Sundvall, M. Laakso, L. Ferrucci, P. Scheet, S. Sanna, M. Uda, Q. Yang, K. L. Lunetta, J. Dupuis, P. I. de Bakker, C. J. O'Donnell, J. C. Chambers, J. S. Kooner, S. Hercberg, P. Meneton, E. G. Lakatta, A. Scuteri, D. Schlessinger, J. Tuomilehto, F. S. Collins, L. Groop, D. Altshuler, R. Collins, G. M. Lathrop, O. Melander, V. Salomaa, L. Peltonen, M. Orho-Melander, J. M. Ordovas, M. Boehnke, G. R. Abecasis, K. L. Mohlke, and

L. A. Cupples. Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat. Genet.*, 41(1):56–65, Jan 2009.

V. Kipnis, R. J. Carroll, L. S. Freedman, and L. Li. Implications of a new dietary measurement error model for estimation of relative risk: application to four calibration studies. *Am. J. Epidemiol.*, 150(6):642–651, Sep 1999.

V. Kipnis, D. Midthune, L. S. Freedman, S. Bingham, A. Schatzkin, A. Subar, and R. J. Carroll. Empirical evidence of correlated biases in dietary assessment instruments and its implications. *Am. J. Epidemiol.*, 153(4):394–403, Feb 2001.

V. Kipnis, D. Midthune, L. Freedman, S. Bingham, N. E. Day, E. Riboli, P. Ferrari, and R. J. Carroll. Bias in dietary-report instruments and its implications for nutritional epidemiology. *Public Health Nutr*, 5(6A):915–923, Dec 2002.

J. S. Kooner, J. C. Chambers, C. A. Aguilar-Salinas, D. A. Hinds, C. L. Hyde, G. R. Warnes, F. J. Gomez Perez, K. A. Frazer, P. Elliott, J. Scott, P. M. Milos, D. R. Cox, and J. F. Thompson. Genome-wide scan identifies variation in MLXIPL associated with plasma triglycerides. *Nat. Genet.*, 40(2):149–151, Feb 2008.

C. Lamina, S. Coassin, T. Illig, and F. Kronenberg. Look beyond one's own nose: combination of information from publicly available sources reveals an association of GATA4 polymorphisms with plasma triglycerides. *Atherosclerosis*, 219(2):698–703, Dec 2011.

M. B. Lanktree and R. A. Hegele. Gene-gene and gene-environment interactions: new insights into the prevention, detection and management of coronary artery disease. *Genome Med*, 1(2):28, 2009.

Y. Li, C. Willer, S. Sanna, and G. Abecasis. Genotype imputation. *Annu Rev Genomics Hum Genet*, 10:387–406, 2009.

Y. Li, C. J. Willer, J. Ding, P. Scheet, and G. R. Abecasis. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.*, 34(8):816–834, Dec 2010.

J. Marchini and B. Howie. Comparing algorithms for genotype imputation. *Am. J. Hum. Genet.*, 83(4):535–539, Oct 2008.

J. Marchini and B. Howie. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.*, 11(7):499–511, Jul 2010.

J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.*, 39(7):906–913, Jul 2007.

M. Miller, N. J. Stone, C. Ballantyne, V. Bittner, M. H. Criqui, H. N. Ginsberg, A. C. Goldberg, W. J. Howard, M. S. Jacobson, P. M. Kris-Etherton, T. A. Lennie, M. Levi, T. Mazzone, and S. Pennathur. Triglycerides and cardiovascular disease: a scientific statement from the American Heart Association. *Circulation*, 123(20): 2292–2333, May 2011.

R. Nielsen. Genomics: In search of rare human variants. *Nature*, 467(7319):1050–1051, Oct 2010.

C. Sabatti, S. K. Service, A. L. Hartikainen, A. Pouta, S. Ripatti, J. Brodsky, C. G. Jones, N. A. Zaitlen, T. Varilo, M. Kaakinen, U. Sovio, A. Ruokonen, J. Laitinen, E. Jakkula, L. Coin, C. Hoggart, A. Collins, H. Turunen, S. Gabriel, P. Elliot, M. I. McCarthy, M. J. Daly, M. R. Jarvelin, N. B. Freimer, and L. Peltonen. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat. Genet.*, 41(1):35–46, Jan 2009.

R. Saxena, B. F. Voight, V. Lyssenko, N. P. Burtt, P. I. de Bakker, H. Chen, J. J. Roix, S. Kathiresan, J. N. Hirschhorn, M. J. Daly, T. E. Hughes, L. Groop, D. Alt-

shuler, P. Almgren, J. C. Florez, J. Meyer, K. Ardlie, K. Bengtsson Bostrom, B. Isomaa, G. Lettre, U. Lindblad, H. N. Lyon, O. Melander, C. Newton-Cheh, P. Nilsson, M. Orho-Melander, L. Rastam, E. K. Speliotes, M. R. Taskinen, T. Tuomi, C. Guiducci, A. Berglund, J. Carlson, L. Gianniny, R. Hackett, L. Hall, J. Holmkvist, E. Laurila, M. Sjogren, M. Sterner, A. Surti, M. Svensson, M. Svensson, R. Tewhey, B. Blumenstiel, M. Parkin, M. Defelice, R. Barry, W. Brodeur, J. Camarata, N. Chia, M. Fava, J. Gibbons, B. Handsaker, C. Healy, K. Nguyen, C. Gates, C. Sougnez, D. Gage, M. Nizzari, S. B. Gabriel, G. W. Chirn, Q. Ma, H. Parikh, D. Richardson, D. Ricke, and S. Purcell. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*, 316(5829): 1331–1336, Jun 2007.

C. H. Tam, R. C. Ma, W. Y. So, Y. Wang, V. K. Lam, S. Germer, M. Martin, J. C. Chan, and M. C. Ng. Interaction effect of genetic polymorphisms in glucokinase (GCK) and glucokinase regulatory protein (GCKR) on metabolic traits in healthy Chinese adults and adolescents. *Diabetes*, 58(3):765–769, Mar 2009.

A. Tan, J. Sun, N. Xia, X. Qin, Y. Hu, S. Zhang, S. Tao, Y. Gao, X. Yang, H. Zhang, S. T. Kim, T. Peng, X. Lin, L. Li, L. Mo, Z. Liang, D. Shi, Z. Huang, X. Huang, M. Liu, Q. Ding, J. M. Trent, S. L. Zheng, Z. Mo, and J. Xu. A genome-wide association and gene-environment interaction study for serum triglycerides levels in a healthy Chinese male population. *Hum. Mol. Genet.*, 21(7):1658–1664, Apr 2012.

The National Heart, Lung, and Blood Institute (NHLBI). SHARe: SNP Health Association Resource, September 2011. URL `http://www.nhlbi.nih.gov/resources/geneticsgenomics/programs/share.htm`.

The National Heart, Lung, and Blood Institute (NHLBI) SHARe Project. NHLBI SNP Health Association Resource (SHARe) Projects, genome-wide association

studies in NHLBI cohorts, August 2010. URL `http://www.ncbi.nlm.nih.gov/bioproject/51455`.

G. A. Thorisson, A. V. Smith, L. Krishnan, and L. D. Stein. The International HapMap Project Web site. *Genome Res.*, 15(11):1592–1593, Nov 2005.

Sally W. Thurston, Donna Spiegelman, and David Ruppert. Equivalence of regression calibration methods in main study/external validation study designs. *Journal of Statistical Planning and Inference*, 113(2):527 – 539, 2003.

Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):pp. 1–25, 1982.

X. Xin, S. R. Srinivasan, W. Chen, E. Boerwinkle, and G. S. Berenson. Interaction effect of Serine447Stop variant of the lipoprotein lipase gene and C-514T variant of the hepatic lipase gene on serum triglyceride levels in young adults: the Bogalusa Heart Study. *Metab. Clin. Exp.*, 52(10):1337–1342, Oct 2003.

J. Zheng, Y. Li, G. R. Abecasis, and P. Scheet. A comparison of approaches to account for uncertainty in analysis of imputed genotypes. *Genet. Epidemiol.*, 35 (2):102–110, Feb 2011.

Appendix A

SUPPLEMENTARY FIGURES: NATURE OF THE SNP DOSAGE DATA

This Appendix includes the supplementary figures discussed in Section 4.1. The first four plots show genotype dosage versus imputed dosage across the 23 SNPs of interest, using HapMap 1+2 imputation (first pair of plots) and 1000 Genomes imputation (second pair of plots). The last pair of plots show HapMap1+2 imputed dosage versus 1000 Genomes imputed dosage.

Figure A.1: CARe genotypes versus HapMap 1+2 imputation

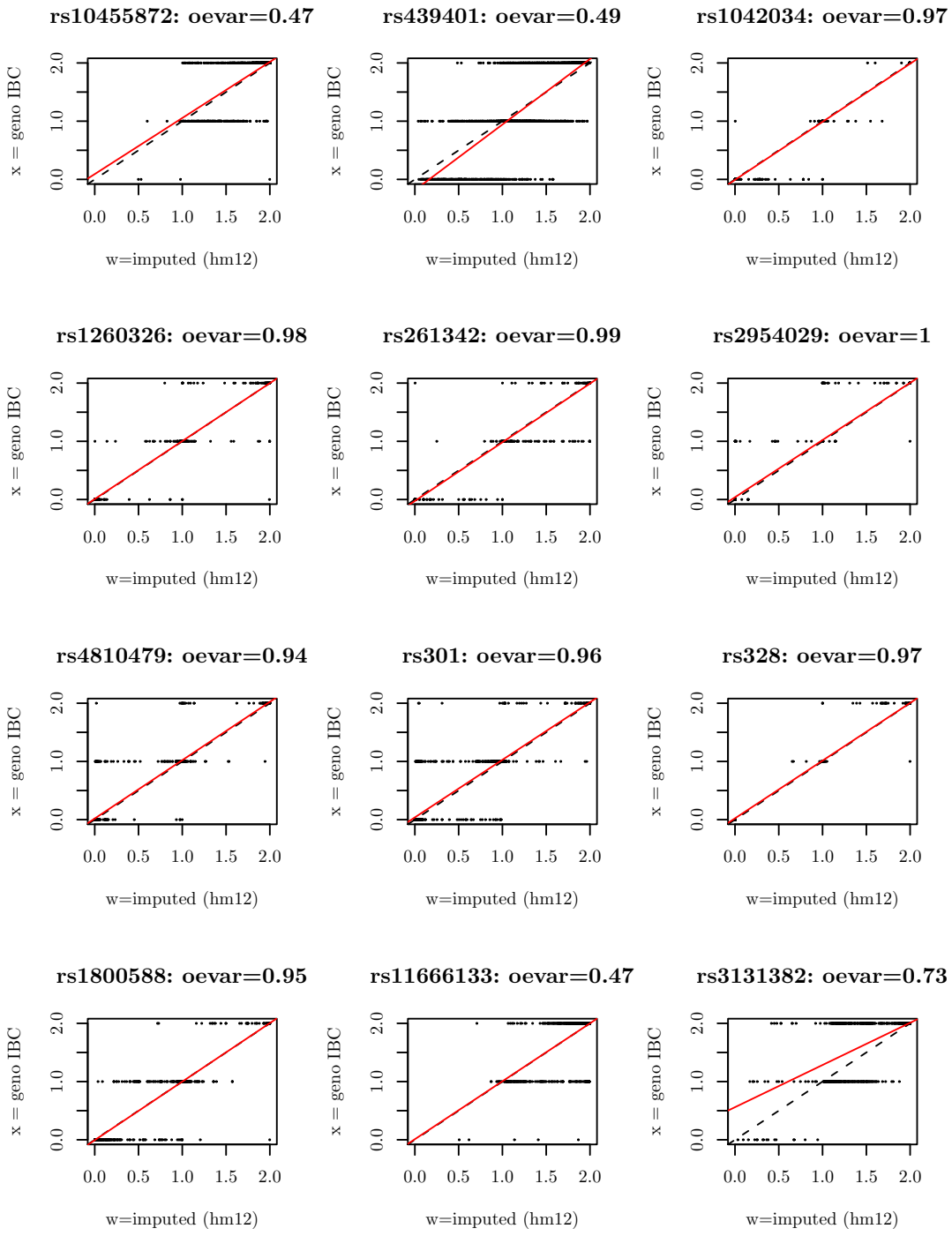


Figure A.2: CARe genotypes versus HapMap 1+2 imputation (continued)

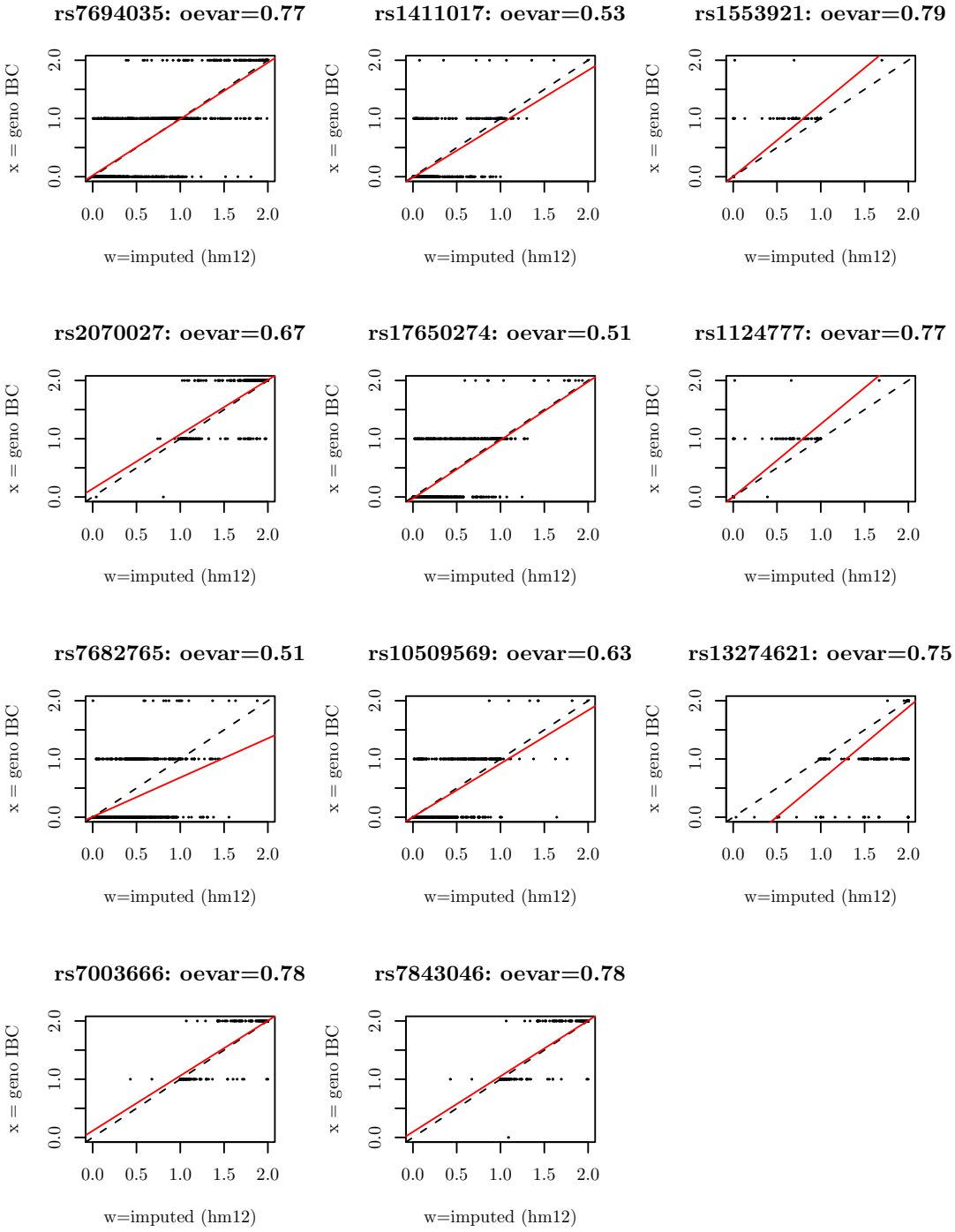


Figure A.3: CARe genotypes versus 1000 Genomes imputation

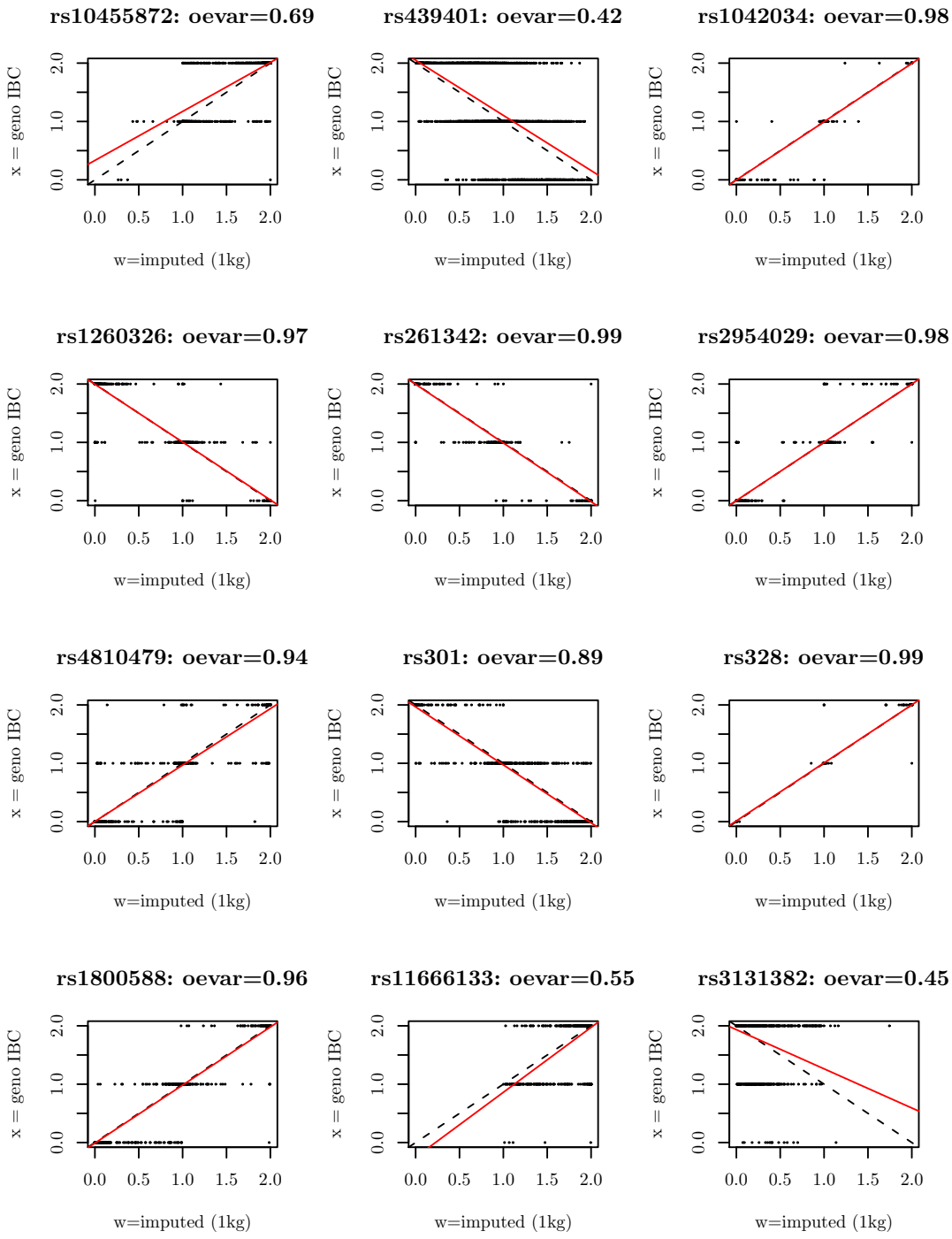


Figure A.4: CARe genotypes versus 1000 Genomes imputation (continued)

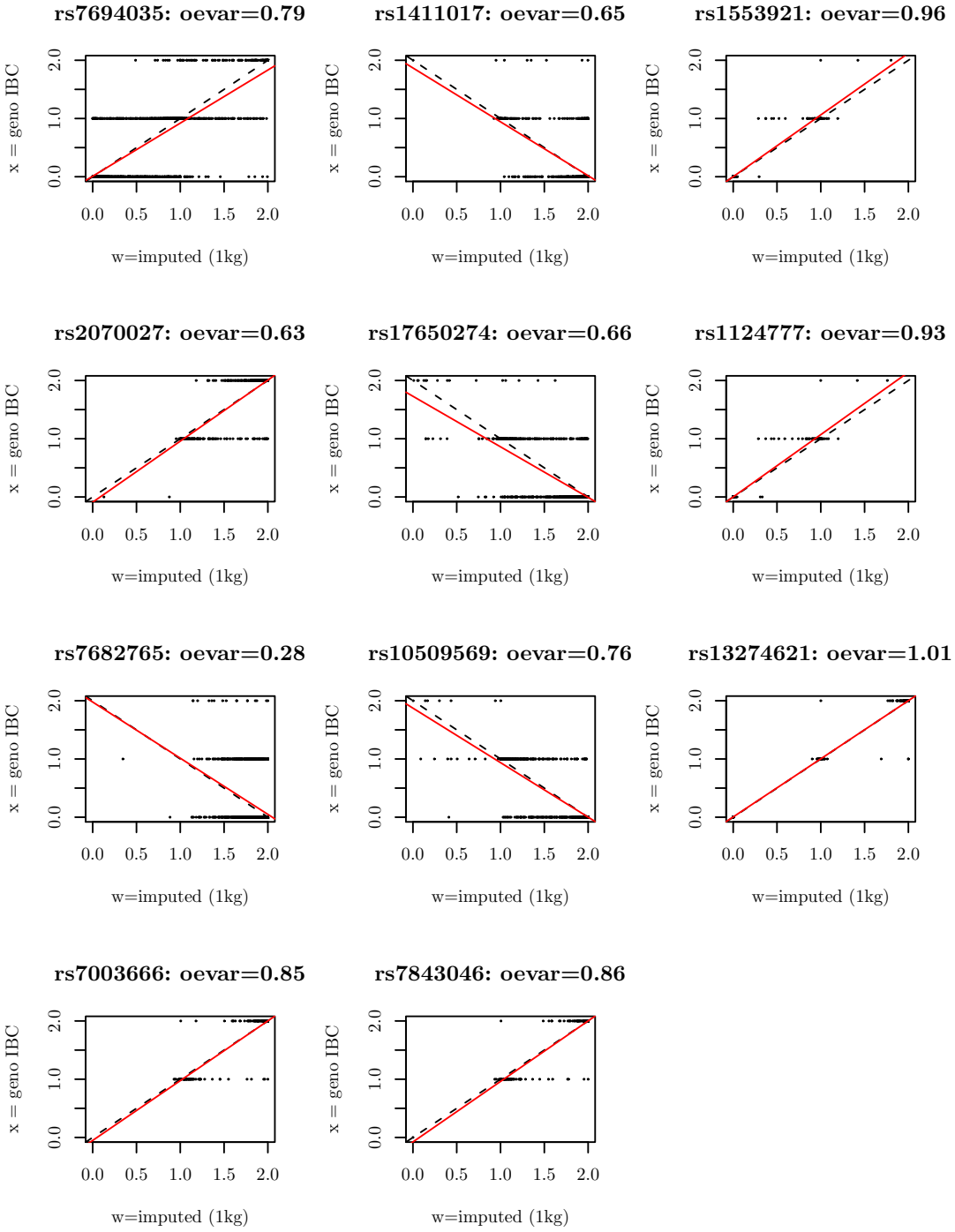


Figure A.5: HapMap 1+2 versus 1000 Genomes imputation

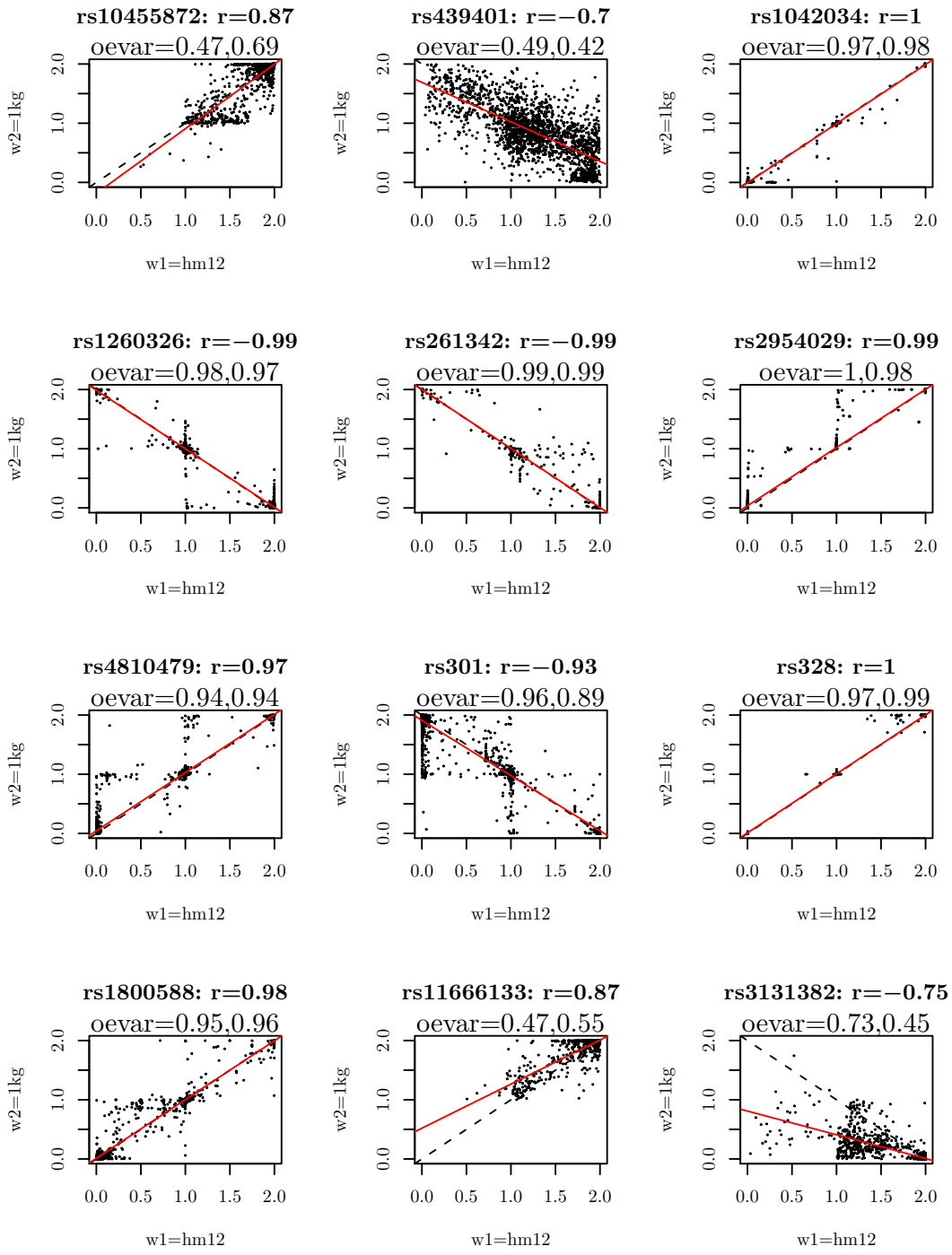
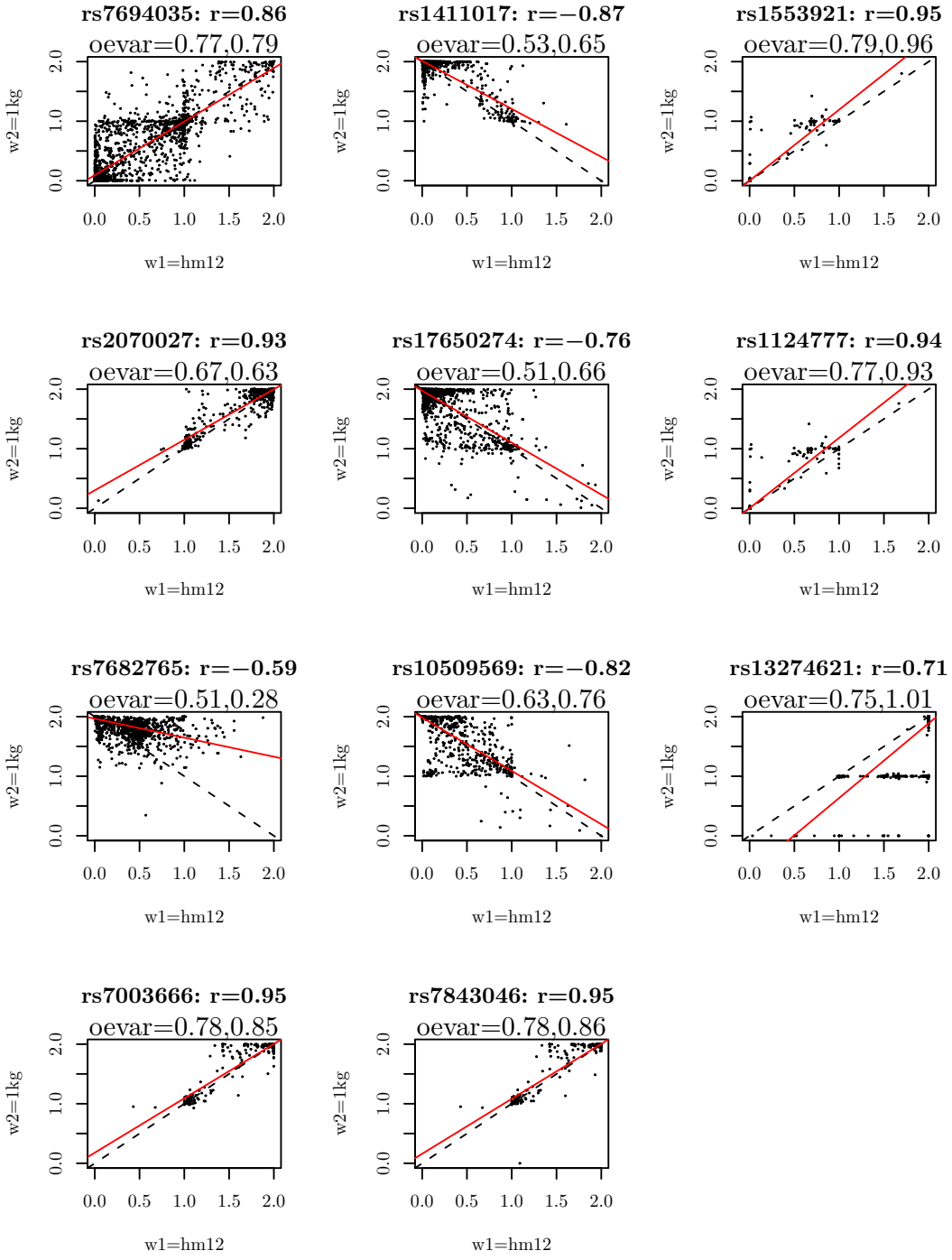


Figure A.6: HapMap 1+2 versus 1000 Genomes imputation (continued)



Appendix B

SINGLE SNP MODEL REGRESSION OUTPUT

This Appendix includes the supplementary table discussed in Section 4.2 and presents regression results using the known data from the CARe genotyping (x), as well as uncorrected estimates from both sources of imputation (w_1 and w_2), compared with validation-based regression calibration ($\hat{x_1}$ and $\hat{x_2}$) for each of the 23 SNPs.

Regression output is presented where $p = 40$, 50, or 90% of the participants have known genotypes (remaining $(1 - p)$% use the calibrated dosage). The subscript s denotes results based on sandwich standard error estimates, m denotes model-based standard error estimates. Sandwich-based results were very similar to bootstrap-based results (not shown). We consider a level of significance of 0.05 because each of these 23 SNPs was previously reported as associated with triglyceride level (Aulchenko et al., 2009; Chasman et al., 2008; Kathiresan et al., 2008, 2009; Kooner et al., 2008; Lamina et al., 2011; Sabatti et al., 2009; Saxena et al., 2007; Tan et al., 2012).

This table shows results of one of the random iterations tabulated in Table 4.2 which presents the total number of SNPs with significant associations detected (out of 23) for the analyses of known genotypes as well as uncorrected and corrected values.

Table B.1: Impact of ME correction in single SNP models

rs10455872	Estimate	Std. Error	95% CI or TI		Pr(>| t |)
x_m	3.330	5.559	−7.565	14.226	0.549
x_s	3.330	5.083	−6.633	13.294	0.512
w_{1m}	−2.444	8.585	−19.270	14.382	0.776
w_{1s}	−2.444	12.032	−26.027	21.139	0.839
$\hat{x}_{1s\cdot 40}$	−3.406	9.913	−22.837	16.024	0.731
$\hat{x}_{1s\cdot 50}$	−1.605	8.947	−19.141	15.932	0.858
$\hat{x}_{1s\cdot 90}$	3.852	5.129	−6.201	13.906	0.453
w_{2m}	−1.239	6.839	−14.643	12.166	0.856
w_{2s}	−1.239	11.180	−23.152	20.675	0.912
$\hat{x}_{2s\cdot 40}$	−4.348	11.553	−26.991	18.296	0.707
$\hat{x}_{2s\cdot 50}$	−3.216	10.558	−23.909	17.476	0.761
$\hat{x}_{2s\cdot 90}$	3.579	5.100	−6.416	13.575	0.483

Table B.1 (continued . . .)

rs439401	Estimate	Std. Error	95% CI or TI		Pr(>| t |)
x_m	10.303	2.942	4.536	16.070	$4.6 \cdot 10^{-4}$
x_s	10.303	2.925	4.570	16.035	$4.3 \cdot 10^{-4}$
w_{1m}	7.558	4.212	−0.697	15.813	0.073
w_{1s}	7.558	3.854	0.004	15.112	0.050
$\hat{x}_{1s\cdot 40}$	8.291	2.988	2.434	14.148	0.006
$\hat{x}_{1s\cdot 50}$	8.440	3.431	1.715	15.165	0.014
$\hat{x}_{1s\cdot 90}$	10.719	3.003	4.834	16.604	$3.6 \cdot 10^{-4}$
w_{2m}	4.208	4.439	−4.492	12.908	0.343
w_{2s}	4.208	3.571	−2.791	11.207	0.239
$\hat{x}_{2s\cdot 40}$	7.940	2.950	2.158	13.722	0.007
$\hat{x}_{2s\cdot 50}$	8.225	3.824	0.731	15.719	0.031
$\hat{x}_{2s\cdot 90}$	10.878	3.072	4.856	16.900	$4.0 \cdot 10^{-4}$
rs1042034	Estimate	Std. Error	95% CI or TI		Pr(>| t |)
x_m	−0.863	3.472	−7.667	5.942	0.804
x_s	−0.863	3.404	−7.535	5.810	0.800
w_{1m}	−0.711	3.489	−7.549	6.126	0.838
w_{1s}	−0.711	3.425	−7.424	6.001	0.835
$\hat{x}_{1s\cdot 40}$	−0.838	3.402	−7.505	5.830	0.805
$\hat{x}_{1s\cdot 50}$	−0.601	3.407	−7.279	6.076	0.860
$\hat{x}_{1s\cdot 90}$	−0.834	3.408	−7.513	5.844	0.807
w_{2m}	−0.611	3.483	−7.439	6.216	0.861
w_{2s}	−0.611	3.418	−7.311	6.088	0.858
$\hat{x}_{2s\cdot 40}$	−0.810	3.405	−7.483	5.863	0.812
$\hat{x}_{2s\cdot 50}$	−0.635	3.405	−7.308	6.038	0.852
$\hat{x}_{2s\cdot 90}$	−0.841	3.407	−7.519	5.837	0.805

Table B.1 (continued . . .)

rs1260326	Estimate	Std. Error	95% CI or TI		Pr(>| t |)
x_m	−6.183	2.877	−11.821	−0.544	0.032
x_s	−6.183	3.261	−12.575	0.209	0.058
w_{1m}	−6.208	2.907	−11.905	−0.510	0.033
w_{1s}	−6.208	3.309	−12.694	0.279	0.061
$\hat{x}_{1s\cdot 40}$	−6.119	3.293	−12.574	0.336	0.063
$\hat{x}_{1s\cdot 50}$	−6.162	3.272	−12.574	0.251	0.060
$\hat{x}_{1s\cdot 90}$	−6.264	3.270	−12.673	0.145	0.055
w_{2m}	−5.666	2.913	−11.375	0.044	0.052
w_{2s}	−5.666	3.364	−12.260	0.928	0.092
$\hat{x}_{2s\cdot 40}$	−5.861	3.315	−12.358	0.636	0.077
$\hat{x}_{2s\cdot 50}$	−5.728	3.311	−12.218	0.761	0.084
$\hat{x}_{2s\cdot 90}$	−6.275	3.268	−12.682	0.131	0.055
rs261342	Estimate	Std. Error	95% CI or TI		Pr(>| t |)
x_m	−6.807	3.372	−13.416	−0.197	0.044
x_s	−6.807	3.319	−13.312	−0.302	0.040
w_{1m}	−6.662	3.462	−13.448	0.124	0.054
w_{1s}	−6.662	3.369	−13.266	−0.058	0.048
$\hat{x}_{1s\cdot 40}$	−6.522	3.344	−13.077	0.033	0.051
$\hat{x}_{1s\cdot 50}$	−6.905	3.282	−13.337	−0.472	0.035
$\hat{x}_{1s\cdot 90}$	−6.690	3.345	−13.246	−0.134	0.046
w_{2m}	−7.211	3.433	−13.940	−0.482	0.036
w_{2s}	−7.211	3.407	−13.889	−0.533	0.034
$\hat{x}_{2s\cdot 40}$	−6.663	3.343	−13.216	−0.110	0.046
$\hat{x}_{2s\cdot 50}$	−7.141	3.361	−13.729	−0.553	0.034
$\hat{x}_{2s\cdot 90}$	−6.649	3.336	−13.187	−0.111	0.046

Table B.1 (continued . . .)

rs2954029	Estimate	Std. Error	95% CI or TI		Pr(>| t |)
x_m	3.527	2.879	−2.116	9.170	0.221
x_s	3.527	2.434	−1.243	8.298	0.147
w_{1m}	3.605	2.868	−2.017	9.227	0.209
w_{1s}	3.605	2.379	−1.058	8.268	0.130
$\hat{x}_{1s\cdot 40}$	3.313	2.445	−1.478	8.104	0.175
$\hat{x}_{1s\cdot 50}$	3.878	2.398	−0.821	8.577	0.106
$\hat{x}_{1s\cdot 90}$	3.578	2.435	−1.195	8.351	0.142
w_{2m}	3.549	2.893	−2.122	9.220	0.220
w_{2s}	3.549	2.428	−1.210	8.308	0.144
$\hat{x}_{2s\cdot 40}$	3.618	2.447	−1.178	8.414	0.139
$\hat{x}_{2s\cdot 50}$	3.488	2.420	−1.255	8.232	0.149
$\hat{x}_{2s\cdot 90}$	3.594	2.437	−1.182	8.371	0.140
rs4810479	Estimate	Std. Error	95% CI or TI		Pr(>| t |)
x_m	0.605	3.221	−5.708	6.918	0.851
x_s	0.605	3.100	−5.471	6.681	0.845
w_{1m}	0.391	3.338	−6.151	6.933	0.907
w_{1s}	0.391	3.059	−5.605	6.386	0.898
$\hat{x}_{1s\cdot 40}$	−0.246	3.191	−6.501	6.009	0.939
$\hat{x}_{1s\cdot 50}$	0.451	3.124	−5.672	6.575	0.885
$\hat{x}_{1s\cdot 90}$	0.465	3.126	−5.662	6.593	0.882
w_{2m}	2.700	3.279	−3.726	9.126	0.410
w_{2s}	2.700	3.148	−3.470	8.870	0.391
$\hat{x}_{2s\cdot 40}$	0.135	3.235	−6.205	6.475	0.967
$\hat{x}_{2s\cdot 50}$	0.964	3.196	−5.301	7.229	0.763
$\hat{x}_{2s\cdot 90}$	0.719	3.108	−5.373	6.810	0.817

Table B.1 (continued . . .)

rs301	Estimate	Std. Error	95% CI or TI		Pr(>| t |)
x_m	−16.277	3.217	−22.581	−9.972	$4.2 \cdot 10^{-7}$
x_s	−16.277	2.929	−22.017	−10.536	$2.7 \cdot 10^{-8}$
w_{1m}	−14.982	3.393	−21.631	−8.332	$1.0 \cdot 10^{-5}$
w_{1s}	−14.982	3.102	−21.062	−8.901	$1.4 \cdot 10^{-6}$
$\hat{x}_{1s\cdot 40}$	−15.452	3.051	−21.433	−9.472	$4.1 \cdot 10^{-7}$
$\hat{x}_{1s\cdot 50}$	−15.608	2.980	−21.449	−9.768	$1.6 \cdot 10^{-7}$
$\hat{x}_{1s\cdot 90}$	−16.104	2.943	−21.872	−10.336	$4.5 \cdot 10^{-8}$
w_{2m}	−16.329	3.397	−22.987	−9.671	$1.5 \cdot 10^{-6}$
w_{2s}	−16.329	3.222	−22.644	−10.014	$4.0 \cdot 10^{-7}$
$\hat{x}_{2s\cdot 40}$	−17.184	3.154	−23.367	−11.002	$5.1 \cdot 10^{-8}$
$\hat{x}_{2s\cdot 50}$	−16.334	3.100	−22.411	−10.257	$1.4 \cdot 10^{-7}$
$\hat{x}_{2s\cdot 90}$	−16.302	2.934	−22.053	−10.552	$2.8 \cdot 10^{-8}$
rs328	Estimate	Std. Error	95% CI or TI		Pr(>| t |)
x_m	17.689	4.575	8.722	26.657	$1.1 \cdot 10^{-4}$
x_s	17.689	3.477	10.874	24.505	$3.6 \cdot 10^{-7}$
w_{1m}	17.796	4.563	8.853	26.739	$9.6 \cdot 10^{-5}$
w_{1s}	17.796	3.491	10.954	24.638	$3.4 \cdot 10^{-7}$
$\hat{x}_{1s\cdot 40}$	17.790	3.502	10.926	24.654	$3.8 \cdot 10^{-7}$
$\hat{x}_{1s\cdot 50}$	17.805	3.514	10.917	24.694	$4.0 \cdot 10^{-7}$
$\hat{x}_{1s\cdot 90}$	17.703	3.480	10.882	24.525	$3.6 \cdot 10^{-7}$
w_{2m}	17.682	4.562	8.740	26.624	$1.1 \cdot 10^{-4}$
w_{2s}	17.682	3.481	10.860	24.504	$3.8 \cdot 10^{-7}$
$\hat{x}_{2s\cdot 40}$	17.623	3.476	10.810	24.436	$4.0 \cdot 10^{-7}$
$\hat{x}_{2s\cdot 50}$	17.632	3.500	10.772	24.492	$4.7 \cdot 10^{-7}$
$\hat{x}_{2s\cdot 90}$	17.691	3.479	10.872	24.510	$3.7 \cdot 10^{-7}$

Table B.1 (continued . . .)

rs1800588	Estimate	Std. Error	95% CI or TI		Pr(>| t |)
x_m	8.184	3.495	1.333	15.034	0.019
x_s	8.184	3.377	1.565	14.802	0.015
w_{1m}	7.667	3.555	0.698	14.635	0.031
w_{1s}	7.667	3.378	1.045	14.288	0.023
$\hat{x}_{1s\cdot 40}$	7.734	3.407	1.055	14.412	0.023
$\hat{x}_{1s\cdot 50}$	7.144	3.312	0.651	13.636	0.031
$\hat{x}_{1s\cdot 90}$	7.981	3.381	1.354	14.607	0.018
w_{2m}	8.313	3.503	1.448	15.179	0.018
w_{2s}	8.313	3.397	1.655	14.972	0.014
$\hat{x}_{2s\cdot 40}$	8.109	3.422	1.401	14.817	0.018
$\hat{x}_{2s\cdot 50}$	8.099	3.354	1.525	14.672	0.016
$\hat{x}_{2s\cdot 90}$	8.177	3.387	1.539	14.816	0.016
rs11666133	Estimate	Std. Error	95% CI or TI		Pr(>| t |)
x_m	17.997	6.716	4.834	31.160	0.007
x_s	17.997	4.451	9.274	26.720	$5.3 \cdot 10^{-5}$
w_{1m}	12.353	9.717	−6.691	31.398	0.204
w_{1s}	12.353	9.263	−5.803	30.510	0.182
$\hat{x}_{1s\cdot 40}$	11.111	7.200	−3.001	25.224	0.123
$\hat{x}_{1s\cdot 50}$	19.804	4.733	10.527	29.081	$2.9 \cdot 10^{-5}$
$\hat{x}_{1s\cdot 90}$	17.323	4.510	8.483	26.163	$1.2 \cdot 10^{-4}$
w_{2m}	19.300	11.346	−2.938	41.537	0.089
w_{2s}	19.300	10.842	−1.950	40.550	0.075
$\hat{x}_{2s\cdot 40}$	14.507	7.269	0.260	28.753	0.046
$\hat{x}_{2s\cdot 50}$	20.493	4.700	11.281	29.704	$1.3 \cdot 10^{-5}$
$\hat{x}_{2s\cdot 90}$	17.150	4.525	8.280	26.020	$1.5 \cdot 10^{-4}$

Table B.1 (continued . . .)

rs3131382	Estimate	Std. Error	95% CI or TI		Pr(>| t |)
x_m	15.667	5.789	4.321	27.013	0.007
x_s	15.667	3.957	7.911	23.422	$7.5 \cdot 10^{-5}$
w_{1m}	14.810	5.766	3.509	26.112	0.010
w_{1s}	14.810	4.451	6.087	23.534	0.001
$\hat{x}_{1s\cdot 40}$	16.860	5.334	6.404	27.315	0.002
$\hat{x}_{1s\cdot 50}$	18.582	4.668	9.433	27.731	$6.9 \cdot 10^{-5}$
$\hat{x}_{1s\cdot 90}$	16.355	4.187	8.148	24.561	$9.4 \cdot 10^{-5}$
w_{2m}	9.779	10.813	−11.414	30.972	0.366
w_{2s}	9.779	8.817	−7.502	27.060	0.267
$\hat{x}_{2s\cdot 40}$	16.242	6.512	3.479	29.005	0.013
$\hat{x}_{2s\cdot 50}$	20.569	5.151	10.472	30.665	$6.5 \cdot 10^{-5}$
$\hat{x}_{2s\cdot 90}$	16.809	4.167	8.642	24.977	$5.5 \cdot 10^{-5}$
rs7694035	Estimate	Std. Error	95% CI or TI		Pr(>| t |)
x_m	12.498	3.290	6.049	18.946	$1.5 \cdot 10^{-4}$
x_s	12.498	3.244	6.139	18.856	$1.2 \cdot 10^{-4}$
w_{1m}	11.050	3.784	3.634	18.465	0.004
w_{1s}	11.050	3.949	3.309	18.790	0.005
$\hat{x}_{1s\cdot 40}$	9.499	3.644	2.357	16.640	0.009
$\hat{x}_{1s\cdot 50}$	11.151	3.624	4.048	18.254	0.002
$\hat{x}_{1s\cdot 90}$	12.608	3.285	6.169	19.047	$1.2 \cdot 10^{-4}$
w_{2m}	12.472	3.633	5.351	19.594	0.001
w_{2s}	12.472	3.213	6.174	18.771	$1.0 \cdot 10^{-4}$
$\hat{x}_{2s\cdot 40}$	11.799	3.194	5.539	18.058	$2.2 \cdot 10^{-4}$
$\hat{x}_{2s\cdot 50}$	12.746	3.587	5.716	19.776	$3.8 \cdot 10^{-4}$
$\hat{x}_{2s\cdot 90}$	12.965	3.315	6.467	19.463	$9.2 \cdot 10^{-5}$

Table B.1 (continued . . .)

rs1411017	Estimate	Std. Error	95% CI or TI		Pr(>| t |)
x_m	−16.411	6.774	−29.688	−3.134	0.015
x_s	−16.411	4.392	−25.019	−7.803	$1.9 \cdot 10^{-4}$
w_{1m}	−15.611	8.532	−32.335	1.112	0.067
w_{1s}	−15.611	5.865	−27.108	−4.115	0.008
$\hat{x}_{1s\cdot 40}$	−15.536	5.144	−25.617	−5.454	0.003
$\hat{x}_{1s\cdot 50}$	−20.445	5.200	−30.636	−10.253	$8.4 \cdot 10^{-5}$
$\hat{x}_{1s\cdot 90}$	−15.371	4.550	−24.290	−6.453	0.001
w_{2m}	−17.165	9.188	−35.174	0.844	0.062
w_{2s}	−17.165	6.213	−29.341	−4.988	0.006
$\hat{x}_{2s\cdot 40}$	−17.978	5.135	−28.042	−7.914	$4.6 \cdot 10^{-4}$
$\hat{x}_{2s\cdot 50}$	−22.704	5.242	−32.978	−12.430	$1.5 \cdot 10^{-5}$
$\hat{x}_{2s\cdot 90}$	−14.833	4.524	−23.699	−5.966	0.001
rs1553921	Estimate	Std. Error	95% CI or TI		Pr(>| t |)
x_m	−21.257	9.790	−40.445	−2.068	0.030
x_s	−21.257	5.872	−32.766	−9.747	$2.9 \cdot 10^{-4}$
w_{1m}	−24.590	13.333	−50.723	1.543	0.065
w_{1s}	−24.590	8.310	−40.877	−8.303	0.003
$\hat{x}_{1s\cdot 40}$	−21.203	6.311	−33.572	−8.834	0.001
$\hat{x}_{1s\cdot 50}$	−21.547	6.311	−33.918	−9.177	0.001
$\hat{x}_{1s\cdot 90}$	−21.985	5.988	−33.721	−10.249	$2.4 \cdot 10^{-4}$
w_{2m}	−22.741	10.614	−43.544	−1.937	0.032
w_{2s}	−22.741	6.575	−35.629	−9.853	0.001
$\hat{x}_{2s\cdot 40}$	−21.312	6.069	−33.207	−9.418	$4.5 \cdot 10^{-4}$
$\hat{x}_{2s\cdot 50}$	−21.427	6.010	−33.207	−9.648	$3.6 \cdot 10^{-4}$
$\hat{x}_{2s\cdot 90}$	−21.632	5.964	−33.322	−9.942	$2.9 \cdot 10^{-4}$

Table B.1 (continued . . .)

rs2070027	Estimate	Std. Error	95% CI or TI		Pr(>| t |)
x_m	19.304	8.044	3.537	35.070	0.016
x_s	19.304	5.388	8.743	29.864	$3.4 \cdot 10^{-4}$
w_{1m}	14.007	8.934	−3.502	31.517	0.117
w_{1s}	14.007	6.879	0.525	27.490	0.042
$\hat{x}_{1s\cdot 40}$	18.086	6.533	5.281	30.891	0.006
$\hat{x}_{1s\cdot 50}$	15.240	6.309	2.875	27.605	0.016
$\hat{x}_{1s\cdot 90}$	19.064	5.411	8.458	29.670	$4.3 \cdot 10^{-4}$
w_{2m}	14.925	9.837	−4.357	34.206	0.129
w_{2s}	14.925	7.232	0.750	29.099	0.039
$\hat{x}_{2s\cdot 40}$	17.079	6.114	5.096	29.061	0.005
$\hat{x}_{2s\cdot 50}$	15.227	6.026	3.417	27.037	0.011
$\hat{x}_{2s\cdot 90}$	17.561	5.393	6.991	28.132	0.001
rs17650274	Estimate	Std. Error	95% CI or TI		Pr(>| t |)
x_m	−13.691	5.410	−24.293	−3.088	0.011
x_s	−13.691	3.825	−21.188	−6.194	$3.4 \cdot 10^{-4}$
w_{1m}	−16.702	7.105	−30.628	−2.777	0.019
w_{1s}	−16.702	5.493	−27.468	−5.937	0.002
$\hat{x}_{1s\cdot 40}$	−19.341	4.519	−28.199	−10.484	$1.9 \cdot 10^{-5}$
$\hat{x}_{1s\cdot 50}$	−13.807	4.571	−22.766	−4.849	0.003
$\hat{x}_{1s\cdot 90}$	−14.326	3.961	−22.089	−6.563	$3.0 \cdot 10^{-4}$
w_{2m}	−16.628	6.203	−28.786	−4.471	0.007
w_{2s}	−16.628	4.494	−25.437	−7.820	$2.2 \cdot 10^{-4}$
$\hat{x}_{2s\cdot 40}$	−20.455	4.462	−29.201	−11.708	$4.6 \cdot 10^{-6}$
$\hat{x}_{2s\cdot 50}$	−17.389	4.304	−25.824	−8.953	$5.3 \cdot 10^{-5}$
$\hat{x}_{2s\cdot 90}$	−14.258	3.954	−22.008	−6.509	$3.1 \cdot 10^{-4}$

Table B.1 (continued . . .)

rs1124777	Estimate	Std. Error	95% CI or TI		Pr(>| t |)
x_m	−21.017	9.857	−40.337	−1.697	0.033
x_s	−21.017	5.930	−32.639	−9.395	$3.9 \cdot 10^{-4}$
w_{1m}	−24.834	13.529	−51.350	1.682	0.067
w_{1s}	−24.834	8.458	−41.412	−8.255	0.003
$\hat{x}_{1s\cdot 40}$	−19.843	6.360	−32.308	−7.378	0.002
$\hat{x}_{1s\cdot 50}$	−20.080	6.442	−32.706	−7.454	0.002
$\hat{x}_{1s\cdot 90}$	−20.850	5.988	−32.587	−9.113	$5.0 \cdot 10^{-4}$
w_{2m}	−23.183	10.773	−44.299	−2.067	0.032
w_{2s}	−23.183	6.636	−36.190	−10.176	$4.8 \cdot 10^{-4}$
$\hat{x}_{2s\cdot 40}$	−21.299	6.107	−33.270	−9.329	$4.9 \cdot 10^{-4}$
$\hat{x}_{2s\cdot 50}$	−21.372	6.148	−33.422	−9.321	0.001
$\hat{x}_{2s\cdot 90}$	−21.068	5.935	−32.702	−9.435	$3.9 \cdot 10^{-4}$
rs7682765	Estimate	Std. Error	95% CI or TI		Pr(>| t |)
x_m	−13.162	5.396	−23.738	−2.587	0.015
x_s	−13.162	3.719	−20.452	−5.873	$4.0 \cdot 10^{-4}$
w_{1m}	−8.161	6.670	−21.233	4.912	0.221
w_{1s}	−8.161	5.540	−19.020	2.698	0.141
$\hat{x}_{1s\cdot 40}$	−9.252	5.614	−20.256	1.752	0.099
$\hat{x}_{1s\cdot 50}$	−13.375	5.041	−23.255	−3.496	0.008
$\hat{x}_{1s\cdot 90}$	−14.848	3.900	−22.493	−7.204	$1.4 \cdot 10^{-4}$
w_{2m}	11.717	12.408	−12.603	36.037	0.345
w_{2s}	11.717	10.334	−8.537	31.971	0.257
$\hat{x}_{2s\cdot 40}$	−3.109	5.824	−14.525	8.307	0.594
$\hat{x}_{2s\cdot 50}$	−7.139	5.382	−17.687	3.410	0.185
$\hat{x}_{2s\cdot 90}$	−15.327	3.937	−23.043	−7.612	$9.9 \cdot 10^{-5}$

Table B.1 (continued . . .)

rs10509569	Estimate	Std. Error	95% CI or TI		Pr(>| t |)
x_m	−15.159	5.834	−26.594	−3.723	0.009
x_s	−15.159	4.290	−23.568	−6.750	$4.1 \cdot 10^{-4}$
w_{1m}	−8.632	7.085	−22.519	5.254	0.223
w_{1s}	−8.632	5.659	−19.725	2.460	0.127
$\hat{x}_{1s\cdot 40}$	−18.265	5.324	−28.700	−7.830	0.001
$\hat{x}_{1s\cdot 50}$	−9.600	5.166	−19.727	0.526	0.063
$\hat{x}_{1s\cdot 90}$	−15.498	4.367	−24.057	−6.939	$3.9 \cdot 10^{-4}$
w_{2m}	−18.679	6.467	−31.355	−6.003	0.004
w_{2s}	−18.679	4.991	−28.461	−8.897	$1.8 \cdot 10^{-4}$
$\hat{x}_{2s\cdot 40}$	−19.774	4.682	−28.951	−10.597	$2.4 \cdot 10^{-5}$
$\hat{x}_{2s\cdot 50}$	−15.909	4.976	−25.662	−6.157	0.001
$\hat{x}_{2s\cdot 90}$	−15.938	4.298	−24.362	−7.515	$2.1 \cdot 10^{-4}$
rs13274621	Estimate	Std. Error	95% CI or TI		Pr(>| t |)
x_m	7.775	4.435	−0.919	16.468	0.080
x_s	7.775	3.658	0.605	14.945	0.034
w_{1m}	1.422	7.836	−13.936	16.780	0.856
w_{1s}	1.422	7.569	−13.414	16.258	0.851
$\hat{x}_{1s\cdot 40}$	5.974	4.768	−3.372	15.319	0.210
$\hat{x}_{1s\cdot 50}$	6.820	4.911	−2.805	16.444	0.165
$\hat{x}_{1s\cdot 90}$	6.329	3.697	−0.916	13.575	0.087
w_{2m}	7.639	4.441	−1.065	16.342	0.086
w_{2s}	7.639	3.662	0.460	14.817	0.037
$\hat{x}_{2s\cdot 40}$	7.665	3.651	0.509	14.820	0.036
$\hat{x}_{2s\cdot 50}$	7.709	3.670	0.516	14.901	0.036
$\hat{x}_{2s\cdot 90}$	7.789	3.655	0.624	14.953	0.033

Table B.1 (continued . . .)

rs7003666	Estimate	Std. Error	95% CI or TI		Pr(>| t |)
x_m	12.489	10.062	−7.233	32.211	0.215
x_s	12.489	7.314	−1.846	26.824	0.088
w_{1m}	10.026	10.492	−10.538	30.591	0.339
w_{1s}	10.026	8.224	−6.093	26.145	0.223
$\hat{x}_{1s\cdot 40}$	12.664	7.851	−2.725	28.053	0.107
$\hat{x}_{1s\cdot 50}$	12.923	7.876	−2.514	28.359	0.101
$\hat{x}_{1s\cdot 90}$	12.872	7.330	−1.496	27.240	0.079
w_{2m}	10.520	10.940	−10.923	31.963	0.336
w_{2s}	10.520	8.447	−6.037	27.077	0.213
$\hat{x}_{2s\cdot 40}$	12.244	7.543	−2.540	27.027	0.105
$\hat{x}_{2s\cdot 50}$	12.263	7.663	−2.756	27.282	0.110
$\hat{x}_{2s\cdot 90}$	12.480	7.332	−1.890	26.850	0.089
rs7843046	Estimate	Std. Error	95% CI or TI		Pr(>| t |)
x_m	13.100	9.929	−6.361	32.561	0.187
x_s	13.100	7.244	−1.098	27.299	0.071
w_{1m}	9.961	10.482	−10.584	30.505	0.342
w_{1s}	9.961	8.219	−6.148	26.069	0.226
$\hat{x}_{1s\cdot 40}$	10.405	8.407	−6.072	26.882	0.216
$\hat{x}_{1s\cdot 50}$	10.862	7.480	−3.799	25.524	0.146
$\hat{x}_{1s\cdot 90}$	13.102	7.350	−1.304	27.508	0.075
w_{2m}	12.980	10.749	−8.089	34.048	0.227
w_{2s}	12.980	8.230	−3.151	29.111	0.115
$\hat{x}_{2s\cdot 40}$	13.168	7.888	−2.292	28.629	0.095
$\hat{x}_{2s\cdot 50}$	13.261	7.609	−1.653	28.174	0.081
$\hat{x}_{2s\cdot 90}$	12.813	7.303	−1.502	27.127	0.079

Appendix C

ANNOTATED R CODE

This Appendix includes R code implementing the proposed measurement error correction described in this thesis. Section C.1 includes annotated R code which implements each of the steps of validation-based regression calibration for a single SNP model. The function's output displays results from uncorrected and corrected approaches as well as the gold standard. Section C.2 details our comparison of validation-based regression calibration with an interaction of two imputed SNPs (using sandwich estimates of standard errors).

C.1 Comparing uncorrected and corrected analysis of single SNP models with the gold standard

```
# Includes comparison of model-based, bootstrapped, and sandwich SEs

library(Hmisc)

valid.regcal <- function(dataset, snpid, ii, B=1000, percent.known)
{
  n  <- dim(dataset)[1]   # number of individuals
  y  <- dataset$trig1     # define phenotype
  z1 <- dataset$age1c     # define covariates z1-z7 (perfectly measured)
  z2 <- dataset$gender1
  z3 <- dataset$site4
  z4 <- dataset$site5
  z5 <- dataset$site6
  z6 <- dataset$site7
  z7 <- dataset$site8

  # define x=genotyped (no ME),
  # w1=HM1+2 imputed (with ME), w2=1000G imputed (with ME)
  x  <- eval(substitute(dataset$variable, list(variable = as.name(paste(snpid[ii],"_C", sep="")))))
```

```
w1 <- eval(substitute(dataset$variable, list(variable = as.name(paste(snpid[ii],"_S", sep="")))))
w2 <- eval(substitute(dataset$variable, list(variable = as.name(paste(snpid[ii],"_S2",sep="")))))

# if coef of w is negative, fix strand flip in w's
if(summary(lm(x~w2))$coeff[2,1]<0)  { w2 <- 2 - w2 }
if(summary(lm(x~w1))$coeff[2,1]<0)  { w1 <- 2 - w1 }

# standard OLS regression for predictors: perfectly measured x,  mismeasured w1, w2
lmperfect  <- lm(y ~ x   + z1+z2+z3+z4+z5+z6+z7)
lmnaive1   <- lm(y ~ w1  + z1+z2+z3+z4+z5+z6+z7)
lmnaive2   <- lm(y ~ w2  + z1+z2+z3+z4+z5+z6+z7)

# now suppose we know x for a subset of percent.known out of n=2026
# get random subset, percent.known of data (same random subset is used for HM1+2 and 1000G)
# r contains the indices of the sample
# is.element(x, y) # length x
sample.n <- round(n*percent.known)   # number of individuals in subset
r        <- sample(1:n, sample.n, replace=F)
known.x  <- is.element(seq(1:n), r)

# motivate correction by comparing regression in x subset only
lmperfect2 <- lm(y ~ x  + z1+z2+z3+z4+z5+z6+z7, subset=known.x)

# correction equation using only x in subset
# predict for all n, known or not, but x.hat stores combo of fitted and known values
# HapMap 1+2
lmrc1        <- lm(x ~ w1  + z1+z2+z3+z4+z5+z6+z7, subset=known.x)
x.fitted1    <- predict(lmrc1, list(w1,z1,z2,z3,z4,z5,z6,z7))
x.hat1       <- ifelse(known.x, x, x.fitted1)

# 1000 Genomes
lmrc2        <- lm(x ~ w2  + z1+z2+z3+z4+z5+z6+z7, subset=known.x)
x.fitted2    <- predict(lmrc2, list(w2,z1,z2,z3,z4,z5,z6,z7))
x.hat2       <- ifelse(known.x, x, x.fitted2)

# regression calibration: OLS regression on corrected values, x.hat; SEs not valid (see below)
lmrc.final1  <- lm(y ~ x.hat1 + z1+z2+z3+z4+z5+z6+z7)
lmrc.final2  <- lm(y ~ x.hat2 + z1+z2+z3+z4+z5+z6+z7)

# regression calibration: same as above but uses corrected SEs (sandwich estimates)
lmfitter<-function(geno)
{
```

```
  # create design matrix Xfit and drop individuals with NA values
  dropitems <- is.na(geno)
  Z     <- as.matrix(cbind(z1,z2,z3,z4,z5,z6,z7))
  # add last column to Z for the intercept
  Z     <- cbind(Z, array(1, c(dim(Z)[1],1)))
  Xfit  <- as.matrix(cbind(geno,Z))
  Xfit  <- Xfit[!dropitems,]
  y     <- y[!dropitems]

  p     <- NCOL(Z)+1
  n     <- NROW(Xfit)
  model <- lm.fit(Xfit, y)
  beta  <- coef(model)[1]
  r     <- resid(model)

  # get sandwich estimates
  # dim(model$qr$qr) = n x p = 760 x p
  A     <- chol2inv(model$qr$qr[1:p,1:p])
  ABA   <- crossprod((Xfit*r)%*%A)
  se    <- sqrt(ABA[1,1])

  # get p-value
  pval  <- 2*pnorm(-abs(beta)/se)

  # return coefficient estimate, sandwich SE, and p-value
  return(c(beta=beta, se=se, pval=pval))
}

# obtain 5 sandwich estimates: x,w1,w2,xh1,xh2
# returns c(beta, se, pval)
lm.xs  <- lmfitter(x)
lm.w1  <- lmfitter(w1)
lm.w2  <- lmfitter(w2)
lm.xh1 <- lmfitter(x.hat1)
lm.xh2 <- lmfitter(x.hat2)

# obtain bootstrap confidence intervals

# subfunction: repeat original analysis in ONE random resample (only saves the coeff estimate)
boot.rc <- function(n)
{
  r <- sample(1:n,n, replace=T)
```

```
  # r denotes the bootstrap resample (same resamples used for HM1+2 and 1000G)
  # code below just repeats the original analysis for this resampled data
  # HapMap 1+2
  lmrc.r1       <- lm(x[r] ~ w1[r] +z1[r]+z2[r]+z3[r]+z4[r]+z5[r]+z6[r]+z7[r], subset=known.x[r])
  x.fitted.r1   <- predict(lmrc.r1, list(w1[r],z1[r],z2[r],z3[r],z4[r],z5[r],z6[r],z7[r]))
  x.hat.r1      <- ifelse(known.x[r], x[r], x.fitted1[r])
                                    # used x.fitted1 values not x.fitted.r1 due to instability
  lmrc.final.r1 <- lm(y[r]~x.hat.r1 +z1[r]+z2[r]+z3[r]+z4[r]+z5[r]+z6[r]+z7[r])

  # 1000 Genomes
  lmrc.r2       <- lm(x[r] ~ w2[r] +z1[r]+z2[r]+z3[r]+z4[r]+z5[r]+z6[r]+z7[r], subset=known.x[r])
  x.fitted.r2   <- predict(lmrc.r2, list(w2[r],z1[r],z2[r],z3[r],z4[r],z5[r],z6[r],z7[r]))
  x.hat.r2      <- ifelse(known.x[r], x[r], x.fitted2[r])
                                    # used x.fitted2 values not x.fitted.r2 due to instability
  lmrc.final.r2 <- lm(y[r]~x.hat.r2 + z1[r]+z2[r]+z3[r]+z4[r]+z5[r]+z6[r]+z7[r])

  # only saving coefficients for SNP from each hm12 and 1kg
  c(lmrc.final.r1$coefficients[2], lmrc.final.r2$coefficients[2])
}

# bootstrap confidence interval needs to allow for uncertainty in x.hat
# B = number of bootstrap resamples to take - length of r below
# call above boot.rc subfunction B times, then get means, sd, quantiles for coefficient estimates
boot.tab <- replicate(B, boot.rc(n))
tab  <- t(apply(boot.tab,  1, function(s){c(mean(s),sqrt(var(s)), quantile(s,c(0.025,0.975)) )}))
# boot SEs = tab[1:2,2]
# boot CIs = tab[1:2,3:4]

beta1.hats <- c(summary(lmperfect  )$coefficients[2,1], lm.xs[1],
                summary(lmnaive1   )$coefficients[2,1], lm.w1[1],
                summary(lmrc.final1)$coefficients[2,1], lm.xh1[1],
                summary(lmnaive2   )$coefficients[2,1], lm.w2[1],
                summary(lmrc.final2)$coefficients[2,1], lm.xh2[1])
SEs        <- c(summary(lmperfect  )$coefficients[2,2], lm.xs[2],
                summary(lmnaive1   )$coefficients[2,2], lm.w1[2], tab[1,2], lm.xh1[2],
                summary(lmnaive2   )$coefficients[2,2], lm.w2[2], tab[2,2], lm.xh2[2])

# 95% CIs, dummy placeholder for TI for now, replace later
low <- beta1.hats - 1.96 * SEs
upp <- beta1.hats + 1.96 * SEs

# use dummy placeholder NA for bootstrap pval
```

```
  pval <- c(summary(lmperfect)$coefficients[2,4], lm.xs[3],
            summary(lmnaive1 )$coefficients[2,4], lm.w1[3], NA, lm.xh1[3],
            summary(lmnaive2 )$coefficients[2,4], lm.w2[3], NA, lm.xh2[3])

  # replace using rows 1-2 (xh1,xh2) of tab[] with tolerance interval low, upp
  # (quantiles of bootstrap resamples)
  low[5] <- tab[1,3]
  upp[5] <- tab[1,4]
  low[9] <- tab[2,3]
  upp[9] <- tab[2,4]

  rownames <- c("xm","xs","w1m","w1s","xhat1b","xhat1s","w2m","w2s","xhat2b","xhat2s")
  # colnames <- c("Estimate", "Std. Error", "2.5%", "97.5%", "Pr(>|t|)")

  # this returns 10 rows corresponding to each model for comparison
  table.B.1 <- as.data.frame(round(cbind(beta1.hats, SEs,low,upp,pval), dig=3), row.names=rownames)
  table.B.1
}
```

C.2 Comparing uncorrected and corrected analysis of $g \times g$ interaction models with the gold standard

```
# Returns 5-row block for each model, for each of main effect of SNP A, SNP B, and interaction AxB
# Uses sandwich SE estimates only

library(Hmisc)

valid.regcal.int <- function(dataset, snpid, ii, B=1000, percent.known)
{
  n  <- dim(dataset)[1]   # number of individuals
  y  <- dataset$trig1     # define phenotype
  z1 <- dataset$age1c     # define covariates z1-z7 (perfectly measured)
  z2 <- dataset$gender1
  z3 <- dataset$site4
  z4 <- dataset$site5
  z5 <- dataset$site6
  z6 <- dataset$site7
  z7 <- dataset$site8

  # define x=genotyped (no ME), w1=HM1+2 imputed (with ME), w2=1000G imputed (with ME)
  # 1=hm12, 2=1kg, a=snpA, b=snpB
  xa  <- eval(substitute(dataset$variable,list(variable=as.name(paste(snpid[ii], "_C", sep="")))))
```

```
w1a <- eval(substitute(dataset$variable,list(variable=as.name(paste(snpid[ii], "_S", sep="")))))
w2a <- eval(substitute(dataset$variable,list(variable=as.name(paste(snpid[ii],"_S2", sep="")))))
xb  <- eval(substitute(dataset$variable,list(variable=as.name(paste(snpid[ii+1],"_C",sep="")))))
w1b <- eval(substitute(dataset$variable,list(variable=as.name(paste(snpid[ii+1],"_S",sep="")))))
w2b <- eval(substitute(dataset$variable,list(variable=as.name(paste(snpid[ii+1],"_S2",sep="")))))

# if coef of w is negative, fix strand flip in w's
if(summary(lm(xa~w2a))$coeff[2,1]<0)  { w2a <- 2 - w2a }
if(summary(lm(xb~w2b))$coeff[2,1]<0)  { w2b <- 2 - w2b }
if(summary(lm(xb~w1b))$coeff[2,1]<0)  { w1b <- 2 - w1b }

# duplicate x's for naming simplicity
x1a <- xa
x2a <- xa
x1b <- xb
x2b <- xb

# create interaction terms
x1ax1b <- x1a * x1b
w1aw1b <- w1a * w1b
x2ax2b <- x2a * x2b
w2aw2b <- w2a * w2b

# standard OLS regression for predictors: perfectly measured x,  mismeasured w1, w2
lmperfect1   <- lm(y ~ x1a + x1b + x1ax1b + z1+z2+z3+z4+z5+z6+z7)
lmnaive1     <- lm(y ~ w1a + w1b + w1aw1b + z1+z2+z3+z4+z5+z6+z7)
# lmperfect2 <- lm(y ~ x2a + x2b + x2ax2b + z1+z2+z3+z4+z5+z6+z7)  # same as lmperfect1
lmnaive2     <- lm(y ~ w2a + w2b + w2aw2b + z1+z2+z3+z4+z5+z6+z7)

# now suppose we know x for a subset of percent.known out of n=2026
# get random subset, percent.known of data (same random subset is used for HM1+2 and 1000G)
# r contains the indices of the sample
# is.element(x, y) # length x
sample.n <- round(n*percent.known)
r        <- sample(1:n, sample.n, replace=F)
known.x <- is.element(seq(1:n), r)

# correction equation using only x in subset (and all covariates)
# predict for all n, known or not, but x.hat stores combo of fitted and known values
# HapMap 1+2
lmrc1a  <- lm(x1a ~ w1a + w1b + w1aw1b + z1+z2+z3+z4+z5+z6+z7, subset=known.x)
lmrc1b  <- lm(x1b ~ w1a + w1b + w1aw1b + z1+z2+z3+z4+z5+z6+z7, subset=known.x)
```

```
x.fitted1a  <- predict(lmrc1a, list(w1a,w1b,w1aw1b,z1,z2,z3,z4,z5,z6,z7))
x.fitted1b  <- predict(lmrc1b, list(w1a,w1b,w1aw1b,z1,z2,z3,z4,z5,z6,z7))
x.hat1a     <- ifelse(known.x, x1a, x.fitted1a)
x.hat1b     <- ifelse(known.x, x1b, x.fitted1b)

# 1000 Genomes
lmrc2a  <- lm(x2a ~ w2a + w2b + w2aw2b + z1+z2+z3+z4+z5+z6+z7, subset=known.x)
lmrc2b  <- lm(x2b ~ w2a + w2b + w2aw2b + z1+z2+z3+z4+z5+z6+z7, subset=known.x)
x.fitted2a  <- predict(lmrc2a, list(w2a,w2b,w2aw2b,z1,z2,z3,z4,z5,z6,z7))
x.fitted2b  <- predict(lmrc2b, list(w2a,w2b,w2aw2b,z1,z2,z3,z4,z5,z6,z7))
x.hat2a     <- ifelse(known.x, x2a, x.fitted2a)
x.hat2b     <- ifelse(known.x, x2b, x.fitted2b)

# calibrate on x1 and x2 scale, not x1x2 scale to preserve linear error structure
x.hat1ab   <- x.hat1a * x.hat1b
x.hat2ab   <- x.hat2a * x.hat2b

# regression calibration: OLS regression on corrected values, x.hat; SEs not valid (see below)
lmrc.final1  <- lm(y ~ x.hat1a +  x.hat1b +  x.hat1ab + z1+z2+z3+z4+z5+z6+z7)
lmrc.final2  <- lm(y ~ x.hat2a +  x.hat2b +  x.hat2ab + z1+z2+z3+z4+z5+z6+z7)

# regression calibration: same as above but uses corrected SEs (sandwich estimates)
lmfitter<-function(geno)
{
  # create design matrix Xfit and drop individuals with NA values
  dropitems<-is.na(geno[,1]) | is.na(geno[,2]) | is.na(geno[,3])
  Z      <-as.matrix(cbind(z1,z2,z3,z4,z5,z6,z7))
  # add last column to Z for the intercept
  Z      <- cbind(Z, array(1, c(dim(Z)[1],1)))
  Xfit   <-as.matrix(cbind(geno,Z))
  Xfit   <-Xfit[!dropitems,]
  y      <-y[!dropitems]

  p      <-NCOL(Xfit)
  n      <-NROW(Xfit)
  model  <-lm.fit(Xfit, y)
  beta   <-coef(model)[1:3]
  r      <-resid(model)

  # get sandwich estimates
  # dim(model$qr$qr) = n x p = 760 x p
  A      <-chol2inv(model$qr$qr[1:p,1:p])
```

```
  ABA    <-crossprod((Xfit*r)%*%A)
  se     <-sqrt(diag(ABA[1:3,1:3]))

  # get p-value
  pval   <- 2*pnorm(-abs(beta)/se)

  # return coefficient estimate, sandwich SE, and p-value
  return(cbind(beta=beta, se=se, pval=pval))
}

# obtain 5 sandwich estimates: x,w1,w2,xh1,xh2
# returns c(beta, se, pval)
lm.xs  <- lmfitter(cbind(x1a, x1b, x1ax1b))
lm.w1  <- lmfitter(cbind(w1a, w1b, w1aw1b))
lm.w2  <- lmfitter(cbind(w2a, w2b, w2aw2b))
lm.xh1 <- lmfitter(cbind(x.hat1a, x.hat1b, x.hat1ab))
lm.xh2 <- lmfitter(cbind(x.hat2a, x.hat2b, x.hat2ab))

# obtain bootstrap confidence intervals

# subfunction: repeat original analysis in ONE random resample (only saves the coeff estimate)
boot.rc <- function(n)
{
  r <- sample(1:n,n, replace=T)
  # r denotes the bootstrap resample (same resamples used for HM1+2 and 1000G)
  # code below just repeats the original analysis for this resampled data
  # HapMap 1+2
  lmrc.r1a <- lm(x1a[r] ~ w1a[r]+ w1b[r]+w1aw1b[r] + z1[r]+z2[r]+z3[r]+z4[r]+z5[r]+z6[r]+z7[r],
                                                                          subset=known.x[r])
  lmrc.r1b <- lm(x1b[r] ~ w1a[r]+ w1b[r]+w1aw1b[r] + z1[r]+z2[r]+z3[r]+z4[r]+z5[r]+z6[r]+z7[r],
                                                                          subset=known.x[r])
  x.fitted.r1a <- predict(lmrc.r1a,
                        list(w1a[r],w1b[r],w1aw1b[r],z1[r],z2[r],z3[r],z4[r],z5[r],z6[r],z7[r]))
  x.fitted.r1b <- predict(lmrc.r1b,
                        list(w1a[r],w1b[r],w1aw1b[r],z1[r],z2[r],z3[r],z4[r],z5[r],z6[r],z7[r]))
  x.hat.r1a    <- ifelse(known.x[r], x1a[r], x.fitted1a[r])
                                    # used x.fitted1 values not x.fitted.r1 due to instability
  x.hat.r1b    <- ifelse(known.x[r], x1b[r], x.fitted1b[r])
  x.hat.r1ab   <- x.hat.r1a * x.hat.r1b
  lmrc.final.r1 <- lm(y[r]~x.hat.r1a + x.hat.r1b + x.hat.r1ab
                                              + z1[r]+z2[r]+z3[r]+z4[r]+z5[r]+z6[r]+z7[r])
```

```
  # 1000 Genomes
  lmrc.r2a <- lm(x2a[r] ~ w2a[r]+ w2b[r] + w2aw2b[r] + z1[r]+z2[r]+z3[r]+z4[r]+z5[r]+z6[r]+z7[r],
                                                                          subset=known.x[r])
  lmrc.r2b <- lm(x2b[r] ~ w2a[r]+ w2b[r] + w2aw2b[r] + z1[r]+z2[r]+z3[r]+z4[r]+z5[r]+z6[r]+z7[r],
                                                                          subset=known.x[r])
  x.fitted.r2a <- predict(lmrc.r2a,
                          list(w2a[r],w2b[r],w2aw2b[r],z1[r],z2[r],z3[r],z4[r],z5[r],z6[r],z7[r]))
  x.fitted.r2b <- predict(lmrc.r2b,
                          list(w2a[r],w2b[r],w2aw2b[r],z1[r],z2[r],z3[r],z4[r],z5[r],z6[r],z7[r]))
  x.hat.r2a    <- ifelse(known.x[r], x2a[r], x.fitted2a[r])
                                      # used x.fitted2 values not x.fitted.r2 due to instability
  x.hat.r2b    <- ifelse(known.x[r], x2b[r], x.fitted2b[r])
  x.hat.r2ab   <- x.hat.r2a * x.hat.r2b
  lmrc.final.r2 <- lm(y[r]~x.hat.r2a + x.hat.r2b + x.hat.r2ab
                                                 + z1[r]+z2[r]+z3[r]+z4[r]+z5[r]+z6[r]+z7[r])

  # only saving coefficients for snpA, snpB, snpAB from each hm12 and 1kg
  c(lmrc.final.r1$coefficients[2:4], lmrc.final.r2$coefficients[2:4])
}

# bootstrap confidence interval needs to allow for uncertainty in x.hat
# B = number of bootstrap resamples to take - length of r below
# call above boot.rc subfunction B times, then get means, sd, quantiles for coefficient estimates
boot.tab <- replicate(B, boot.rc(n))
tab  <- t(apply(boot.tab,  1, function(s){c(mean(s),sqrt(var(s)), quantile(s,c(0.025,0.975)) )}))
# boot SEs = tab[1:6,2]
# boot CIs = tab[1:6,3:4]

# note the 5 sandwich models
# xa, w1a, xh1a, w2a, xh2a
beta1.hats  <- c(lm.xs[1,1], lm.w1[1,1], lm.xh1[1,1], lm.w2[1,1], lm.xh2[1,1])
beta2.hats  <- c(lm.xs[2,1], lm.w1[2,1], lm.xh1[2,1], lm.w2[2,1], lm.xh2[2,1])
beta12.hats <- c(lm.xs[3,1], lm.w1[3,1], lm.xh1[3,1], lm.w2[3,1], lm.xh2[3,1])
SEs1        <- c(lm.xs[1,2], lm.w1[1,2], lm.xh1[1,2], lm.w2[1,2], lm.xh2[1,2])
SEs2        <- c(lm.xs[2,2], lm.w1[2,2], lm.xh1[2,2], lm.w2[2,2], lm.xh2[2,2])
SEs12       <- c(lm.xs[3,2], lm.w1[3,2], lm.xh1[3,2], lm.w2[3,2], lm.xh2[3,2])

# 95% CIs
low1  <- beta1.hats  - 1.96 * SEs1
upp1  <- beta1.hats  + 1.96 * SEs1
low2  <- beta2.hats  - 1.96 * SEs2
upp2  <- beta2.hats  + 1.96 * SEs2
```

```
  low12 <- beta12.hats - 1.96 * SEs12
  upp12 <- beta12.hats + 1.96 * SEs12

  # sandwich pval
  pval1        <- c(lm.xs[1,3], lm.w1[1,3], lm.xh1[1,3], lm.w2[1,3], lm.xh2[1,3])
  pval2        <- c(lm.xs[2,3], lm.w1[2,3], lm.xh1[2,3], lm.w2[2,3], lm.xh2[2,3])
  pval12       <- c(lm.xs[3,3], lm.w1[3,3], lm.xh1[3,3], lm.w2[3,3], lm.xh2[3,3])

  rownames1  <- c("x_as","w_1as","hatx_1as","w_2as","hatx_2as")
  rownames2  <- c("x_bs","w_1bs","hatx_1bs","w_2bs","hatx_2bs")
  rownames12 <- c("x_abs","w_1abs","hatx_1abs","w_2abs","hatx_2abs")
  # colnames <- c("Estimate", "Std. Error", "2.5%", "97.5%", "Pr(>|t|)")

  Estimate <- beta1.hats
  SE       <- SEs1
  low      <- low1
  upp      <- upp1
  pval     <- pval1
  table_4_4_a <- as.data.frame(round(cbind(Estimate, SE,low,upp,pval), dig=3), row.names=rownames1)

  Estimate <- beta2.hats
  SE       <- SEs2
  low      <- low2
  upp      <- upp2
  pval     <- pval2
  table_4_4_b <- as.data.frame(round(cbind(Estimate, SE,low,upp,pval), dig=3), row.names=rownames2)

  Estimate <- beta12.hats
  SE       <- SEs12
  low      <- low12
  upp      <- upp12
  pval     <- pval12
  table_4_4_ab <- as.data.frame(round(cbind(Estimate,SE,low,upp,pval),dig=3), row.names=rownames12)

  # Returns 5-row block for each model, for each of main effect of SNP A, SNP B, and int AxB
  table.4.4 <- as.data.frame(rbind(table_4_4_a,table_4_4_b,table_4_4_ab))
  table.4.4
}
```