Missing Data Methods in Mendelian Randomization Studies With Multiple Instruments

Stephen Burgess* , Shaun Seaman, Debbie A. Lawlor, Juan P. Casas, and Simon G. Thompson

* Correspondence to Stephen Burgess, Medical Research Council Biostatistics Unit, Robinson Way, Cambridge CB2 0SR, United Kingdom (e-mail: stephen.burgess@mrc-bsu.cam.ac.uk).

Initially submitted February 17, 2011; accepted for publication June 15, 2011.

Mendelian randomization studies typically have low power. Where there are several valid candidate genetic instruments, precision can be gained by using all the instruments available. However, sporadically missing genetic data can offset this gain. The authors describe 4 Bayesian methods for imputing the missing data based on a missing-at-random assumption: multiple imputations, single nucleotide polymorphism (SNP) imputation, latent variables, and haplotype imputation. These methods are demonstrated in a simulation study and then applied to estimate the causal relation between C-reactive protein and each of fibrinogen and coronary heart disease, based on 3 SNPs in British Women’s Heart and Health Study participants assessed at baseline between May 1999 and June 2000. A complete-case analysis based on all 3 SNPs was found to be more precise than analyses using any 1 SNP alone. Precision is further improved by using any of the 4 proposed missing data methods; the improvement is equivalent to about a 25% increase in sample size. All methods gave similar results, which were apparently not overly sensitive to violation of the missing-at-random assumption. Programming code for the analyses presented is available online.

Bayesian methods; causal inference; imputation; instrumental variables; Mendelian randomization analysis; missing data

Abbreviation: SNP, single nucleotide polymorphism.

Mendelian randomization uses genetic variants as instrumental variables to estimate causal associations of a risk factor on an outcome in situations where there is potential unmeasured confounding or reverse causation (1). We refer to such a risk factor, protective factor, or intermediate phenotype simply as a “phenotype.” A genetic variant can be used as an instrumental variable to divide a population into genotypic subgroups in an analogous way to how randomization divides participants into arms in a randomized controlled trial (2). Under certain assumptions about the genetic variant (G), these subgroups are systematically different in their exposure to the phenotype of interest (X) but not in any possible confounders (U) (3). Any difference in outcome (Y) between the subgroups is causally due to the phenotype (4). The necessary assumptions for the instrumental variable analysis (5) are that G is

1) independent of any possible confounders,
2) associated with the phenotype, and
3) independent of the outcome given the phenotype and confounders.

These assumptions are summarized in the directed acyclic graph (Figure 1) (6).

One difficulty with Mendelian randomization is that, although the instrumental variable estimate is consistent (asymptotically unbiased) for the causal association, its variance is typically much larger than the variance from a standard analysis (i.e., regression of Y on X adjusted for known confounders) (7). This is because the variation in the phenotype explained by the instrumental variable is usually small (8, 9). To test some causal associations, sample sizes of several thousands are needed (10).

A possible solution is to use multiple instrumental variables. Where there are several genetic variants that can be used as instrumental variables and each explains independent variation in the phenotype, the instrumental variable estimate using all of the instruments will have lower variance than the instrumental variable estimate using a subset of the instrumental variables (11, 12). However, a problem arising from including multiple instrumental variables in an analysis is missing

1069
data (13). Sporadically missing genetic data typically arise from difficulty in interpreting the output of genotyping platforms. If the output is not clear, a “missing” result is recorded. Hence, although efficiency will be gained from using multiple instruments, this may be offset in a complete-case analysis because more participants with missing data are omitted.

Rather than omitting participants, we seek to use the structure of the genetic data, in particular the correlation between genetic markers known as linkage disequilibrium, to impute missing data and include all participants in an analysis, acknowledging uncertainty in the imputation. In this paper, we introduce 4 methods for imputing missing data under the missing-at-random assumption; that is, the pattern of missingness in the genotype data does not depend on the values of the missing genetic data but only on the genetic data that are observed (14). We use a Bayesian method that is robust to weak instrument bias (15, 16) and discuss possible modifications if data are missing not at random; that is, missingness depends also on the unobserved missing values. We apply these methods in a simulation study and to real data from the British Women’s Heart and Health Study on the association between C-reactive protein and coronary heart disease. The observational associations between C-reactive protein and fibrinogen and between C-reactive protein and coronary heart disease are both positive but attenuate on adjustment for known confounders. It is thought that the true causal associations are null (17, 18).

METHODS FOR INCORPORATING MISSING DATA

We conduct our analyses in a Bayesian framework as this lends itself naturally to data imputation. We first introduce a Bayesian complete-case analysis method (19) and then 4 methods for imputing data under the missing-at-random assumption that can be incorporated into the Bayesian model to include subjects with missing data.

We assume throughout that all instrumental variables are single nucleotide polymorphisms (SNPs) with 2 possible alleles. We code each SNP as 0, 1, or 2, representing the number of variant alleles. Individuals with 1 variant allele on a SNP are heterozygotes; otherwise they are homozygotes. A per-allele genetic model is presumed for each SNP; another model could be used if considered more appropriate.

Genetic data may be missing for several reasons: An individual may fail to provide a sample for analysis, consent may not be given for genetic testing, the DNA extracted may be of insufficient quality or quantity for analysis, or the reading from a genetic platform may be difficult to interpret and hence a missing result may be recorded. In the first 3 instances, no genetic data would be available for the individual, but they could be included in the analysis. Although they would be informative about the distribution of the phenotype and outcome, they would not generally contribute greatly to the estimation of a causal effect. The focus of this work is on individuals who have missing data for only some SNPs, as these would contribute most to the estimation of the causal effect.

Bayesian model

With continuous outcomes, our method is analogous to the 2-stage least-squares method (3), except that it acknowledges the observational correlation between phenotype and outcome. This means that it is less biased with weak instruments and that the credible intervals have close to nominal coverage (20). For each individual $i$, the phenotype $x_i$ and the continuous outcome $y_i$ are assumed to come from a bivariate normal distribution with correlation $\rho$. We model the average phenotype level ($\zeta_i$) per allele within instruments and additive across instruments ($\beta_{ik} = 0, 1, 2$, $k = 1, \ldots, K$) and the average outcome level ($\eta_i$) as depending linearly on the average phenotype.

$$
\begin{align*}
(x_i, y_i) &\sim N_2 \left( \begin{pmatrix} \zeta_i \\ \eta_i \end{pmatrix}, \begin{pmatrix} \sigma^2_{x} & \rho \sigma_x \sigma_y \\ \rho \sigma_x \sigma_y & \sigma^2_y \end{pmatrix} \right) \\
\zeta_i &\sim \alpha_0 + \sum_{k=1}^{K} \alpha_k g_{ik} \\
\eta_i &\sim \beta_0 + \beta_1 \zeta_i
\end{align*}
$$

(1)

With binary outcomes, we use a Bernoulli distribution for the outcome $y_i = 0$ or 1, with probability of event $\pi_i$ related to the phenotype by a logistic model (21).

$$
\begin{align*}
x_i &\sim N(\zeta_i, \sigma^2_x) \\
y_i &\sim \text{Bernoulli}(\pi_i) \\
\logit(\pi_i) &\sim \beta_0 + \beta_1 \zeta_i
\end{align*}
$$

(2)

In all binary outcome analyses, we make inference on the gene-phenotype association only in individuals without prior history of disease (4).

In each case, the causal parameter of interest is $\beta_1$, the increase in outcome (or log-odds of outcome) per unit increase in the phenotype. We use vague prior distributions on all parameters: In our example, these are normal priors with mean zero and variance 1,000$^2$ for all regression parameters, uniform priors on $[0, 20]$ for standard deviations, and a uniform prior on $[-1, 1]$ for the correlation $\rho$. We use Monte Carlo Markov chain sampling by WinBUGS (22) for all analyses, with at least 50,000 iterations, of which the first 1,000 are discarded as “burn-in.” We assess convergence by running
Missingness in either phenotype or outcome is easily dealt with by the model, as information on \( \xi \) and \( \eta \) is gained from all individuals with data on phenotype and outcome. However, missingness in the instrumental variables is less simple, as it is not clear what the underlying distribution of the genetic parameters is. We present 4 methods for addressing missing genetic data below.

**Multiple imputations method.** We first impute the genetic data multiple times using a genetic software package (we used Beagle (24, 25) in this paper) and incorporate the imputations into the Bayesian model using the WinBUGS dpick function (22) to choose one of the imputed data sets at random in each Markov chain iteration. Beagle imputes genetic data using a hidden Markov model and empirical Bayes methods under a missing-at-random assumption. The dpick function gives a discrete uniform categorical random variable taking integer values such that feedback from the rest of the model to this random variable is not permitted (26), so that the imputed data sets are used equally often on average. Model 1 or model 2 is modified as follows:

\[
m - \text{Discrete uniform}(1, M) \tag{3}
\]

\[
\xi_i = \alpha_0 + \sum_{k=1}^K \alpha_k g_{ikm},
\]

where \( g_{ikm} \) is the number of variant alleles of SNP \( k \) for individual \( i \) in the imputed data set \( m, m = 1, \ldots, M \). When \( M = \infty \), this is equivalent to imputing from the posterior distribution of the genotypes given by the genetic software package without feedback. This is a similar idea to classical multiple imputation but implemented in a Bayesian setting.

In the examples below, we use \( M = 10 \) imputations.

**SNP imputation method.** Instead of using the multiple imputations approach, we can use the posterior probabilities of genotypes given by the same software package for each SNP directly in the Bayesian model. The output from Beagle gives us posterior probabilities \( p_{ijk} \) that SNP \( k \) for individual \( i \) takes value \( j \). We model the number of variant alleles of SNP \( k \) for individual \( i \) as a categorical random variable taking values in \( \{0, 1, 2\} \). Model 1 or model 2 is modified by adding the following:

\[
g_{ik} - \text{Categorical}(p_{0k}, p_{1k}, p_{2k}). \tag{4}
\]

A disadvantage of this method is that it does not account for known correlation between SNPs when imputing multiple SNPs in the same individual. Additionally, in both the multiple and SNP imputation methods, only the genetic data are used to impute missing values. As the phenotype and outcome data contain information about the missing genetic data values, they should also be used in the imputation model (27). However, if the genetic markers are highly correlated and the genetic data do not explain much variation in the phenotype, then we would not expect the bias caused by this omission to be large.

**Multivariate latent variables method.** In this method, we extended our Bayesian model to include the Bayesian model for imputation of correlated SNPs proposed by Lunn et al. (28). Genetic material in humans is arranged in 2 haplotypes, each consisting of combinations of alleles which are inherited together. We use latent vectors \( \psi_{1i} = (\psi_{1i1}, \ldots, \psi_{1iK}) \) and \( \psi_{2i} = (\psi_{2i1}, \ldots, \psi_{2iK}) \) to model each of the haplotypes for an individual \( i \) by a multivariate normal random variable with one component corresponding to each SNP. If \( \psi_{1ik} \) is positive, SNP \( k \) on the first haplotype (numbered arbitrarily) has a variant allele; otherwise not. Hence, the number of variant alleles for SNP \( k \) is \( I(\psi_{1ik} > 0) + I(\psi_{2ik} > 0) \), where \( I(\cdot) \) is an indicator function. We use the WinBUGS function \( \text{dgene.aux} \) to model the number of variant alleles (28). This function describes a discrete distribution on \( \{0, 1, 2\} \) taking 2 arguments. When both arguments are negative, \( \text{dgene.aux} \) is 0 with probability 1; when the arguments have an opposite sign, \( \text{dgene.aux} \) is 1 with probability 1; when both are positive, \( \text{dgene.aux} \) is 2 with probability 1. The function is coded as a probability distribution rather than as a deterministic function for technical reasons: Missing genetic data values are required to be stochastic, rather than deterministic nodes. The latent variables are a convenient way of modeling correlations in discrete distributions with analogy to the underlying biologic structure of the problem.

\[
\psi_{1i} - \mathcal{N}_K(\mu, \Sigma) \tag{5}
\]

\[
\psi_{2i} - \mathcal{N}_K(\mu, \Sigma)
\]

\[
g_{ik} - \text{dgene.aux}(\psi_{1ik}, \psi_{2ik})
\]

The mean \( (\mu) \) and variance-covariance matrix \( (\Sigma) \) of the multivariate normal distribution are given vague priors: multivariate independent normal and inverse Wishart, respectively.

**Haplotype imputation method.** If the variation in the genetic data can be summarized by a small number of haplotypes, then instead of using a SNP-based model of genetic association, we can use a haplotype-based model. If individual \( i \) has haplotypes \( h_{1i} \) and \( h_{2i} \), we have the following:

\[
\xi_i = \gamma_{h_{1i}} + \gamma_{h_{2i}}. \tag{6}
\]

Often, when there is substantial genetic variation, SNPs are chosen to tag haplotypes, and there is a 1:1 correspondence between SNPs and haplotypes. In this case, a per-allele additive SNP-based model is equivalent to this additive haplotype model. When there is uncertainty in haplotype assignment due to missing data, we use the available SNPs to reduce the genetic variation in the data to a set of candidate haplotypes and model each unknown haplotype value by a categorical random variable, with probabilities for each haplotype estimated from the relative proportions of each of the possible haplotypes in the data set. This method is illustrated for a specific data set below.

A disadvantage of this method is that it is very difficult to write a general model that could be used for arbitrary genetic data. A separate imputation model is needed for each genotypic pattern of observed and missing data in the study population. This method is not recommended when there is
an uncertainty in haplotype assignment for individuals with complete data, as the model may lose identifiability.

**Simulation study**

We perform a simulation study to assess the performance of the 4 imputation methods. Three genetic variants (G₁, G₂, G₃) are used as instrumental variables. Following Figure 1, the data are generated from the model:

\[
X = \alpha_1 G_1 + \alpha_2 G_2 + \alpha_3 G_3 + U + E_X
\]

\[
Y = \beta_1 X - 2U + E_Y
\]

\[U, E_X, E_Y \sim N(0, 1)\] independently, (8)

and missing data are introduced by random draws \(R_k\) for SNP \(k\) for each individual, where \(G_k\) is observed if \(R_k = 1\) and missing if \(R_k = 0\). The true causal effect was \(\beta_1 = 1\). Data sets of 1,000 individuals were generated for a range of 5 realistic scenarios:

- **Scenario 1** has \(\mathbb{P}(R_1 = 1) = \mathbb{P}(R_2 = 1) = 1\), \(\mathbb{P}(R_3 = 1) = 0.8\), so that only SNP 3 contains any missingness. SNPs 2 and 3 are taken to be in complete linkage disequilibrium. Minor allele frequencies are all 0.4.

- **Scenario 2** has correlated SNPs tagging 4 haplotypes with frequencies 0.4, 0.3, 0.2, and 0.1. \(R_1, R_2\), and \(R_3\) are independent with \(\mathbb{P}(R_j = 1|G_1, G_2, G_3) = 0.93\).

- **Scenario 3** has the same missingness mechanism as scenario 2, but SNPs are uncorrelated. Minor allele frequencies are 0.4, 0.4, and 0.2.

- **Scenario 4** has the same haplotypes as scenario 2, but \(R_1, R_2, \) and \(R_3\) are independent with \(\mathbb{P}(R_j = 1|G_1, G_2, G_3) = 0.98\) if \(G_j = 0\) or 2 (i.e., homozygous at SNP \(j\)), and \(\mathbb{P}(R_j = 1|G_1, G_2, G_3) = 0.88\) if \(G_j = 1\) (i.e., heterozygous).

- **Scenario 5** has the same missingness mechanism as scenario 4 but the same uncorrelated SNPs as scenario 3.

Parameters of genetic association \((x_1, x_2, x_3)\) were chosen to give an average \(F\) statistic of around 16–20. In each scenario, the complete-case analysis contains on average around 20% fewer individuals than the complete-data analysis because of missingness. Scenarios 1–3 follow the missing-at-random assumption, while scenarios 4 and 5 do not.

Results are given in Table 1 on the basis of 1,000 simulated data sets for each scenario (100 for the SNP imputation method for computational reasons). We note that the haplotype imputation analysis is possible in scenario 1, but results would be as the complete-data analysis, because data would be imputed without uncertainty. The haplotype imputation is not attempted in scenarios 3 and 5 as the SNPs are uncorrelated. All other models, including the latent variables model (which estimates a variance-covariance matrix with near-zero correlation in scenarios 3 and 5), have been applied in each scenario.

The relative efficiency (ratio of Monte Carlo variances) of the complete-case analysis compared with the complete-data analysis is 0.77–0.84, consistent with the 20% decrease in sample size. All of the methods outperform the complete-case analysis in all scenarios, with no evident bias and reduced variance of the causal estimate. In scenario 1, where the missing data can be imputed with minimal uncertainty, each of the imputation methods performs almost identically to the complete-data analysis. The multiple imputations method seems to perform better with uncorrelated SNPs (scenarios 2 and 4) and is outperformed by the haplotype imputation method with correlated SNPs (scenarios 3 and 5). The latent variable approach has spuriously high precision with a relative efficiency greater than 1 in some scenarios, although with apparently correct coverage of the 95% confidence interval, possibly due to lack of convergence. The methods cope well when the missing-at-random assumption is violated, with minimal bias and a gain in precision compared with the complete-case analysis. Further details of the simulation study are provided in Web Appendix 1, the first of 2 Web appendices posted on the *Journal’s* Web site (http://aje.oupjournals.org/).

### Table 1. Mean Estimate of Causal Effect \(\beta_1 = 1\) (Relative Efficiency Compared With Complete-Data Analysis) Across 1,000 Simulated Data Sets for Complete-Data Analysis, Complete-Case Analysis, and 4 Imputation Methods in 5 Scenarios

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Complete-Data Analysis</th>
<th>Complete-Case Analysis</th>
<th>Multiple Imputations Method</th>
<th>SNP Imputation Method*</th>
<th>Latent Variables Method</th>
<th>Haplotype Imputation Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario 1</td>
<td>1.012 (1)</td>
<td>1.019 (0.79)</td>
<td>1.013 (1.00)</td>
<td>1.00 (0.99)</td>
<td>1.005 (1.05)</td>
<td>—*b</td>
</tr>
<tr>
<td>Scenario 2</td>
<td>1.028 (1)</td>
<td>1.031 (0.77)</td>
<td>1.024 (0.82)</td>
<td>1.03 (0.94)</td>
<td>1.012 (0.96)</td>
<td>1.010 (0.91)</td>
</tr>
<tr>
<td>Scenario 3</td>
<td>1.012 (1)</td>
<td>1.017 (0.80)</td>
<td>1.012 (0.90)</td>
<td>1.00 (0.90)</td>
<td>1.001 (1.00)</td>
<td>N/A</td>
</tr>
<tr>
<td>Scenario 4</td>
<td>1.006 (1)</td>
<td>1.009 (0.84)</td>
<td>1.008 (0.85)</td>
<td>1.00 (0.96)</td>
<td>1.002 (0.96)</td>
<td>0.996 (0.92)</td>
</tr>
<tr>
<td>Scenario 5</td>
<td>1.013 (1)</td>
<td>1.018 (0.81)</td>
<td>1.015 (0.97)</td>
<td>0.97 (0.97)</td>
<td>1.008 (1.02)</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Abbreviations: N/A, not attempted; SNP, single nucleotide polymorphism.
* Results for the SNP imputation method are across only 100 simulations for computational reasons, so results are presented to only 2 decimal places.
* * —, results would be exactly as in the complete-data analysis, refer to the text for details.
Table 2. Haplotypes in the C-reactive Protein Gene Region Tagged by 3 Single Nucleotide Polymorphisms Used as Instruments

<table>
<thead>
<tr>
<th>Haplotype</th>
<th>rs1205</th>
<th>rs1130864</th>
<th>rs1800947</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>C</td>
<td>T</td>
<td>G</td>
</tr>
<tr>
<td>2</td>
<td>C</td>
<td>C</td>
<td>G</td>
</tr>
<tr>
<td>3</td>
<td>T</td>
<td>C</td>
<td>G</td>
</tr>
<tr>
<td>4</td>
<td>T</td>
<td>C</td>
<td>C</td>
</tr>
</tbody>
</table>

British Women’s Heart and Health Study

We illustrate our methods by using data from the British Women's Heart and Health Study to assess the impact of using multiple instruments and missing data on Mendelian randomization analyses. We examine the causal effect of C-reactive protein on fibrinogen (continuous outcome) and on coronary heart disease (binary outcome) using 3 SNPs in the C-reactive protein gene region as instrumental variables: rs1205, rs1130864, and rs1800947. These 3 SNPs tag 4 haplotypes (Table 2) that comprise over 99% of the variation in the C-reactive protein gene in European-descent populations (29).

The British Women’s Heart and Health Study is a prospective cohort study of heart disease in British women between the ages of 60 and 79 years. We use cross-sectional baseline data on 3,693 participants who have complete or partial data for C-reactive protein, fibrinogen, and the 3 SNPs. There is missingness in 21% of participants for C-reactive protein, 2.4% for fibrinogen, 10.8% for rs1205, 1.9% for rs1130864, and 2.6% for rs1800947. Genotyping was undertaken by KBioscience (Hoddesdon, Hertfordshire, United Kingdom) on 2 separate occasions for SNP rs1205 and then for SNPs rs1130864 and rs1800947. Table 3 shows the pattern of missingness of SNPs. Although it is unusual to see so many more missing data in one SNP than in another, this may be due to the individual characteristics of that SNP or region of the DNA. A total of 3,188 individuals had data on all 3 SNPs; of these, 12 (0.4%) individuals had a genotype that did not conform to the haplotype patterns of Table 2. C-reactive protein measurements were assessed by using an immunonephelometric high-sensitivity assay (Dade Behring, Marburg, Germany). To rule out reverse causation, only C-reactive protein measurements from nondiseased individuals were considered. Coronary heart disease was defined as nonfatal myocardial infarction (using World Health Organization criteria). We assessed coronary heart disease at baseline, comparing individuals who had had a definite previous myocardial infarction (6.9%) with all other individuals. C-reactive protein was log-transformed throughout. We found that a per-allele model of genetic association was appropriate for each of the SNPs. Each of the SNPs was in Hardy-Weinberg equilibrium. Only participants of European descent were included to ensure homogeneity of the population in question.

Complete-case analyses. We analyzed the British Women’s Heart and Health Study data using each of the 3 SNPs measured as the sole instrumental variable and with all of the SNPs included as instrumental variables. We performed 2 sets of analyses, with the first including all participants with complete data on the instrumental variable in question and the second using the common set of 3,188 participants with measured values for all 3 SNPs. The F statistic in multivariate regression of phenotype on all the instruments is 16.7, indicating little potential bias from weak instruments (20).

Table 4 shows that, considering the data on participants with complete data for each of the SNPs, using all the SNPs as the instrumental variable gives the most precise estimator, with at least 20% reduction in standard error compared with using any of the SNPs individually. However, a substantial proportion of the data has been discarded in the complete-case analysis. If we use only SNP rs1130864 as the instrumental variable, an additional 421 participants can be included in the analysis, resulting in about a 20% reduction in standard error. Although this gain in precision is not uniform across all SNPs, with a slight loss of precision in the causal estimates using SNP rs1800947 as the instrumental variable despite a sample size increase of 396, this analysis motivates us to use methods for incorporating individuals with missing data.

Haplotype-based analysis. For the haplotype imputation method, we note that each of the SNPs available here tags 1 haplotype. This means that the haplotype assignment of an individual with complete genetic data that are consistent with haplotypes 1–4 of Table 2 can be determined without uncertainty. Where there are missing data, we consider the possible haplotype assignments consistent with the 4 haplotypes of Table 2. For example, an individual measured as heterozygous in SNPs rs1205 and rs1800947 (CT and CG) with a missing data value for SNP rs1130864 must have 1 copy of haplotype 4 and 1 copy of either haplotype 1 or 2. An individual measured as homozygous CC in SNP rs1205 and GG in rs1800947 with a missing data value for SNP rs1130864 has 2 haplotypes that must each be either 1 or 2. For each individual, we model the unknown haplotypes using categorical random variables. For example, the variables for these cases would each have a binomial distribution taking value 1 or 2 with probabilities corresponding to the relative proportions of the haplotypes in the population. To estimate the proportions of each haplotype, we assume independence of haplotypes within and between individuals and maximize the likelihood of a multinomial distribution with the correct likelihood contributions from individuals with complete and missing data. These probabilities are used to form the priors for the categorical variables in the Bayesian analysis. The
convergence for the parameters in the multivariate latent variable distribution, even when the number of iterations was substantially increased. However, the distribution of the causal parameter seemed to have converged. The reduction in the standard error for all missing-data methods compared with the complete-case analysis is 8%—12%, corresponding to a 17%—29% increase in sample size, slightly more than the true increase in sample size of 16%. The Monte Carlo standard error, which describes the uncertainty about the value of the causal estimate due to using Monte Carlo Markov chain estimation, is approximately 0.002 for the continuous outcome and 0.01 for the binary outcome.

It is perhaps surprising to find a gain in precision greater than the gain in sample size. However, the increase in sample size within each of the genotypic subgroups, each containing all individuals with a particular genotype, is not uniform. In this case, the individuals with imputed data fall disproportionately into the smaller subgroups. This means that most of the smaller subgroups increase in size by more than 16%, giving rise to a greater-than-expected increase in precision. These results assume that the data are missing at random, meaning that the fact that a data value is missing gives no information about the true value of the data point. In Web Appendix 1, we investigate possible departures from the assumption of missingness at random in genetic data (30) and show that these have only limited impact on the estimate of causal association.

RESULTS UNDER THE MISSING-AT-RANDOM ASSUMPTION

We applied each of the 4 methods described above. Each of the imputation methods gives similar answers, which differ somewhat from the complete-case analysis results in terms of point estimate (Table 5), especially in the binary case. The exception is the latent variable method, which reported poor

Table 5. Estimates (SE) and 95% Confidence Interval of the Causal Effect of Unit Increase in Log-transformed C-reactive Protein on Fibrinogen (µmol/L) and Coronary Heart Disease (β_l) in Complete-Case Analysis (n = 3,188) and in the Entire Study Population (n = 3,693) of the British Women’s Heart and Health Study, 1999–2001, by Different Imputation Methods for Missing Genetic Data

<table>
<thead>
<tr>
<th>Imputation Method</th>
<th>Continuous Outcome: Mean Difference in Fibrinogen (SE)</th>
<th>95% CI</th>
<th>Binary Outcome: Log-Odds Ratio of CHD (SE)</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete-case analysis</td>
<td>−0.102 (0.274)</td>
<td>−0.699, 0.382</td>
<td>0.44 (0.55)</td>
<td>−0.57, 1.59</td>
</tr>
<tr>
<td>Multiple imputations method</td>
<td>−0.088 (0.249)</td>
<td>−0.619, 0.358</td>
<td>0.22 (0.50)</td>
<td>−0.75, 1.25</td>
</tr>
<tr>
<td>SNP imputation method</td>
<td>−0.075 (0.250)</td>
<td>−0.613, 0.369</td>
<td>0.22 (0.49)</td>
<td>−0.73, 1.22</td>
</tr>
<tr>
<td>Latent variable methoda</td>
<td>−0.040 (0.241)</td>
<td>−0.552, 0.401</td>
<td>0.20 (0.48)</td>
<td>−0.72, 1.15</td>
</tr>
<tr>
<td>Haplotype imputation method</td>
<td>−0.061 (0.250)</td>
<td>−0.590, 0.391</td>
<td>0.23 (0.51)</td>
<td>−0.75, 1.25</td>
</tr>
</tbody>
</table>

Abbreviations: CHD, coronary heart disease; CI, confidence interval; SE, standard error; SNP, single nucleotide polymorphism.

a The latent variable results are presented with the caveat that the parameters in the multivariate normal distribution of the latent variables did not converge, although the causal parameter did seem to have converged.

DISCUSSION

In this paper, we have considered the impact of using multiple instruments on instrumental variable analysis. Using multiple instruments has the potential to reduce the variance of the causal estimates, but if there are sporadic missing data, this increase is offset by a decrease in sample size in a complete-case analysis. The missing data methods we have described can be used to include all participants and to gain precision in the analysis under the assumption of missingness at random. Even though this assumption may not be fully valid, the results in our example were not sensitive to departures from this assumption. A further assumption of the imputation methods

<table>
<thead>
<tr>
<th>Imputation Method</th>
<th>Continuous Outcome: Mean Difference in Fibrinogen (SE)</th>
<th>95% CI</th>
<th>Binary Outcome: Log-Odds Ratio of CHD (SE)</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete-case analysis</td>
<td>−0.102 (0.274)</td>
<td>−0.699, 0.382</td>
<td>0.44 (0.55)</td>
<td>−0.57, 1.59</td>
</tr>
<tr>
<td>Multiple imputations method</td>
<td>−0.088 (0.249)</td>
<td>−0.619, 0.358</td>
<td>0.22 (0.50)</td>
<td>−0.75, 1.25</td>
</tr>
<tr>
<td>SNP imputation method</td>
<td>−0.075 (0.250)</td>
<td>−0.613, 0.369</td>
<td>0.22 (0.49)</td>
<td>−0.73, 1.22</td>
</tr>
<tr>
<td>Latent variable methoda</td>
<td>−0.040 (0.241)</td>
<td>−0.552, 0.401</td>
<td>0.20 (0.48)</td>
<td>−0.72, 1.15</td>
</tr>
<tr>
<td>Haplotype imputation method</td>
<td>−0.061 (0.250)</td>
<td>−0.590, 0.391</td>
<td>0.23 (0.51)</td>
<td>−0.75, 1.25</td>
</tr>
</tbody>
</table>

Abbreviations: CHD, coronary heart disease; CI, confidence interval; SE, standard error; SNP, single nucleotide polymorphism.

a The latent variable results are presented with the caveat that the parameters in the multivariate normal distribution of the latent variables did not converge, although the causal parameter did seem to have converged.

Am J Epidemiol. 2011;174(9):1069–1076
is that of Hardy-Weinberg equilibrium. However, violation of Hardy-Weinberg equilibrium is often an indication of a population substructure; if a SNP is not in Hardy-Weinberg equilibrium, this may call into question its use as an instrumental variable for Mendelian randomization in the data set.

Although the haplotype imputation model is the most natural of the methods, relying on only the independent inheritance of haplotypes in the study population, it is not necessarily applicable to all Mendelian randomization studies. A characteristic of this data set is that the SNPs can be summarized as a small number of haplotypes with certainty; haplotype imputation in this data set is the preferred analysis.

Of the 3 general purpose methods for missing data imputation, the latent variable method is the most interpretable in terms of the underlying biology. One concern may be that the impact of the distributional assumptions of the latent variables on the analysis is not clear. There is a danger of lack of convergence or poor mixing in complicated Bayesian models such as this, which resulted in a somewhat different estimate from the other methods in the British Women’s Heart and Health Study example with the continuous outcome, although less difference was observed with the binary outcome.

The SNP imputation and the multiple imputations methods are both easy to implement and based on the same idea. In the multiple imputations method, we rely on sampling from a discrete number of imputations rather than from the entire probability distribution, although the number of imputations could be increased if this were thought to be a problem. A drawback of the SNP imputation model is the assumption of prior independence of the SNPs in the imputation. One problem with these 2 methods is that the genetic data are imputed without using the phenotype and outcome. Although we would expect some attenuation in the causal estimate due to the omission of the phenotype, the results seem fairly similar to those of the haplotype and latent variable models, both of which allow feedback from the phenotype and outcome in the imputation process. In the paper, we used Beagle for genetic imputation; results were similar when other imputation programs, such as fastPHASE (31), were used.

Our recommended preference, where possible, would be to use a haplotype imputation method. If this is not possible, because of uncertainty in haplotype ascertainment, we would suggest using the multiple imputations method, with the latent variable method as a sensitivity analysis for the effect of omitting the phenotype and outcome from the imputation model.

The WinBUGS code for the general purpose multiple imputation methods used is available online (32) and in Web Appendix 2.

ACKNOWLEDGMENTS

Author affiliations: Medical Research Council Biostatistics Unit, University of Cambridge, Cambridge, United Kingdom (Stephen Burgess, Shaun Seaman); Medical Research Council Centre for Causal Analyses in Translational Epidemiology, School of Social and Community Medicine, University of Bristol, Bristol, United Kingdom (Debbie A. Lawlor); Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, United Kingdom (Juan P. Casas); and Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom (Simon G. Thompson).

The authors are supported by the Medical Research Council (grants U.1052.00.001, U.1052.00.006, G0600705, and G0601625) of the United Kingdom. The British Women’s Heart and Health Study is supported by the Department of Health (England) Policy Research Programme and the British Heart Foundation.

The authors would like to thank the British Women’s Heart and Health Study for permission to use their data for this methodological study, general practitioners who supported data collection in the British Women’s Heart and Health Study, and the staff who collected, entered, and cleaned data.

The views expressed in this paper are those of the authors and not necessarily those of any funding body.

Conflict of interest: none declared.

REFERENCES

13. Palmer TM, Lawlor DA, Harbord RM, et al. Using multiple genetic variants as instrumental variables for modifiable risk...


