Imputation of Genotype and Haplotype using *IVE*ware

T. E. Raghunathan (Raghu)

1 Introduction

During a lecture on strategies for the analysis of incomplete data, a student asked whether *IVE*ware can be used to multiply impute missing Genotype for SNPs and Haplotypes associated with those genotypes. The application of *IVE*ware to obtain multiple imputations in the statistical genetics area is new. There are several imputation packages specifically designed for statistical genetics but most of them perform single and mostly deterministic imputations. Thus, imputation using the sequential regression approach and a readily available software deserves some attention. The goal of this approach is to be fully conditional on disease and covariates and incorporate imputation uncertainty through multiple imputations.

Suppose that Y is the binary disease status (1=cases and 0=controls) and $X_1, X_2, \ldots, X_p$ is a collection of p covariates (e.g., age, sex, education, income, blood pressure etc). Let $G_1, G_2, \ldots, G_q$ be the genotypes for q SNPs and the legal values for each Genotype is 0, 1 or 2. Of course, not all Genotypes are observed. It is possible that covariates have some missing values.

A haplotype associated with a genotype, G_j is a decomposition, a pair of latent binary variables, (H_{j1}, H_{j2}), such that $H_{j1} + H_{j2} = G_j$. Obviously, if $G_j = 2$ then $H_{j1} = 1$ and $H_{j2} = 1$ and if $G_j = 0$ then $H_{j1} = 0$ and $H_{j2} = 0$. However, if $G_j = 1$ then there are two possible scenarios $H_{j1} = 1, H_{j2} = 0$ or $H_{j1} = 0, H_{j2} = 1$. In this case, if H_{j1} is imputed then H_{j2} is determined as $1 - H_{j1}$.

The goal is to create a multiply imputed completed-data on disease, covariates and genotypes and haplotypes. The standard multiple imputation analysis can then be performed on the completed-data sets and the point estimates and their sampling variances can be combined using the standard multiple imputation formula.

Let X_{obs}, Y_{obs}, G_{obs} and H_{obs} denote observed set of values and X_{mis}, Y_{mis}, G_{mis} and H_{mis} denote the corresponding missing set of values. Imputations should be drawn from the predictive distribution,

$$Pr(X_{mis}, Y_{mis}, G_{mis}, H_{mis}|X_{obs}, Y_{obs}, G_{obs}, H_{obs})$$

$$= Pr(X_{mis}, Y_{mis}, G_{mis}|X_{obs}, Y_{obs}, G_{obs}, H_{obs})\times$$

$$Pr(H_{mis}|X, Y, G, H_{obs}).$$

Note that G_{obs} and H_{obs} are deterministically related to each other. Specifically, if $G_{obs} = 2$ then $H_{obs} = (1, 1)$ and if $G_{obs} = 0$ then $H_{obs} = (0, 0)$. If $G_{obs} = 1$ then there is no corresponding H_{obs}.

Thus, the above decomposition can be written as

$$Pr(X_{mis}, Y_{mis}, G_{mis}, H_{mis}|X_{obs}, Y_{obs}, G_{obs}, H_{obs})$$

$$= Pr(X_{mis}, Y_{mis}, G_{mis}|X_{obs}, Y_{obs}, G_{obs})\times$$

$$Pr(H_{mis}|X, Y, G, H_{obs}).$$

Accordingly, imputation will be carried out in two stages. First stage imputes missing genotypes. This is a straightforward application of *IVE*ware. The second stage is to reformat the imputed data obtained in the first stage to impute the haplotypes using the standard features of *IVE*ware. This method can also be implemented in MICE (for those using R), ICE (for those using STATA) and SRCware (the stand-alone version of *IVE*ware).

2 Imputation of Genotypes

Consider an example data set (assumed to be stored as a SAS data file, eg1.sas7bdat, in the directory "Mydir") described below with genotypes for 4 SNPs ($RS1, RS2, RS3$ and $RS4$),3 covariates ($Age, Gender$ and SBP) and disease status ($Disease$):

Id	Disease	Age	Gender	SBP	RS1	RS2	RS3	RS4
1	1	35	F	125	1	2	?	0
2	1	42	?	?	2	0	1	2
3	0	38	M	148	?	1	0	?
4	0	?	F	162	0	?	2	1
5	1	42	M	118	?	?	2	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

The following example SAS code will impute missing values in the above data set:

```
libname din "C:\mydir";
%impute(name=eg1,dir=C:\mydir,setup=new);
datain din.eg1;
dataout din.eg1imputed;
continuous age spb;
categorical disease gender rs1 rs2 rs3 rs4;
transfer id;
multiples 5;
seed 125647;
iterations 10;
run;
```

Depending upon the sample size, *IVE*ware features such as MAXPRED or MINRSQD may have to be used to reduce the dimensionality of the covariates in the imputation model. Also, to preserve some gene by environment interactions, interactions between genotype and covariates (or between genotypes for gene-gene interactions) may be included in the imputation model.

3 Imputation of Haplotypes

Once the missing covariates and the genotypes have been imputed, then the data needs to be organized into the haplotypes for the next stage of imputation. Suppose that the imputed data from the first stage is as follows:

Id	Disease	Age	Gender	SBP	RS1	RS2	RS3	RS4
1	1	35	F	125	1	2	1	0
2	1	42	F	132	2	0	1	2
3	0	38	M	148	1	1	0	1
4	0	39	F	162	0	2	2	1
5	1	42	M	118	1	0	2	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

We now split each genotype into haplotypes as shown in the following table

Id	Disease	Age	Gender	SBP	RS1h1	RS1h2	RS2h1	RS2h2	RS3h1	RS3h2	RS4
1	1	35	F	125	?	?	1	1	?	?	0
2	1	42	F	132	1	1	0	0	?	?	1
3	0	38	M	148	?	?	?	?	0	0	?
4	0	39	F	162	0	0	1	1	1	1	?
5	1	42	M	118	?	?	0	0	1	1	?
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Now, if $RS1h1$ is known then $RS1h2$ can be derived. Thus, only one member of the pair of binary variables needs to be imputed and the other can be derived from the corresponding genotype $RS1$. Suppose that the above data set is stored in a file called "eg1haplotype". The following *IVE*ware code accomplishes the imputation of haplotypes:

```
%impute(name=eg1,dir=C:\mydir,setup=new);
datain din.eg1haplotype;
dataout din.eg1haplotypeimp;
continuous age spb;
categorical disease gender rs1h1 rs2h1 rs3h1 rs4h1;
transfer id rs1h2 rs2h2 rs3h2 rs4h2;
multiples 5;
seed 7456321;
iterations 10;
run;
```

Finally, merge "eg1imputed" and "eg1haplotypeimp" and replace the missing $RS1h2$ by $RS1-RS1h1$. This merged data set has all the genotypes and haplotypes imputed using the fully conditional or sequential regression specifications. Repeating these steps for each imputed data from Stage 1, will result in 25 multiply imputed data sets.

4 Incorporating external genotype data

In many practical applications, an external reference genotype data on large number of SNPs may be available for imputation of genotype into the sample. For example, Figure 1, provides a schematic of such a data structure. In this situation, the first stage imputation will be based on the concatenated sample data and external reference data just on the SNPs to impute the genotypes of the larger set of SNPs into the sample. The second stage imputation of haplotypes will be based on the fully imputed sample data.

From the design perspective, it may be better to mount a GWAS study by sampling cases and controls from the sample data and obtaining the genotype on larger set of SNPs and then use *IVE*ware to impute the rest. This design is illustrated in the schematic Figure 2.

As described earlier, the application of *IVE*ware to the statistical genetics is new and needs to be thoroughly investigated and compared to other existing approaches. Some actual and simulated data sets may be used to evaluate the multiple imputation inferences of association parameters or regression models etc.

Figure 1: A schematic display of genetics data to be imputed

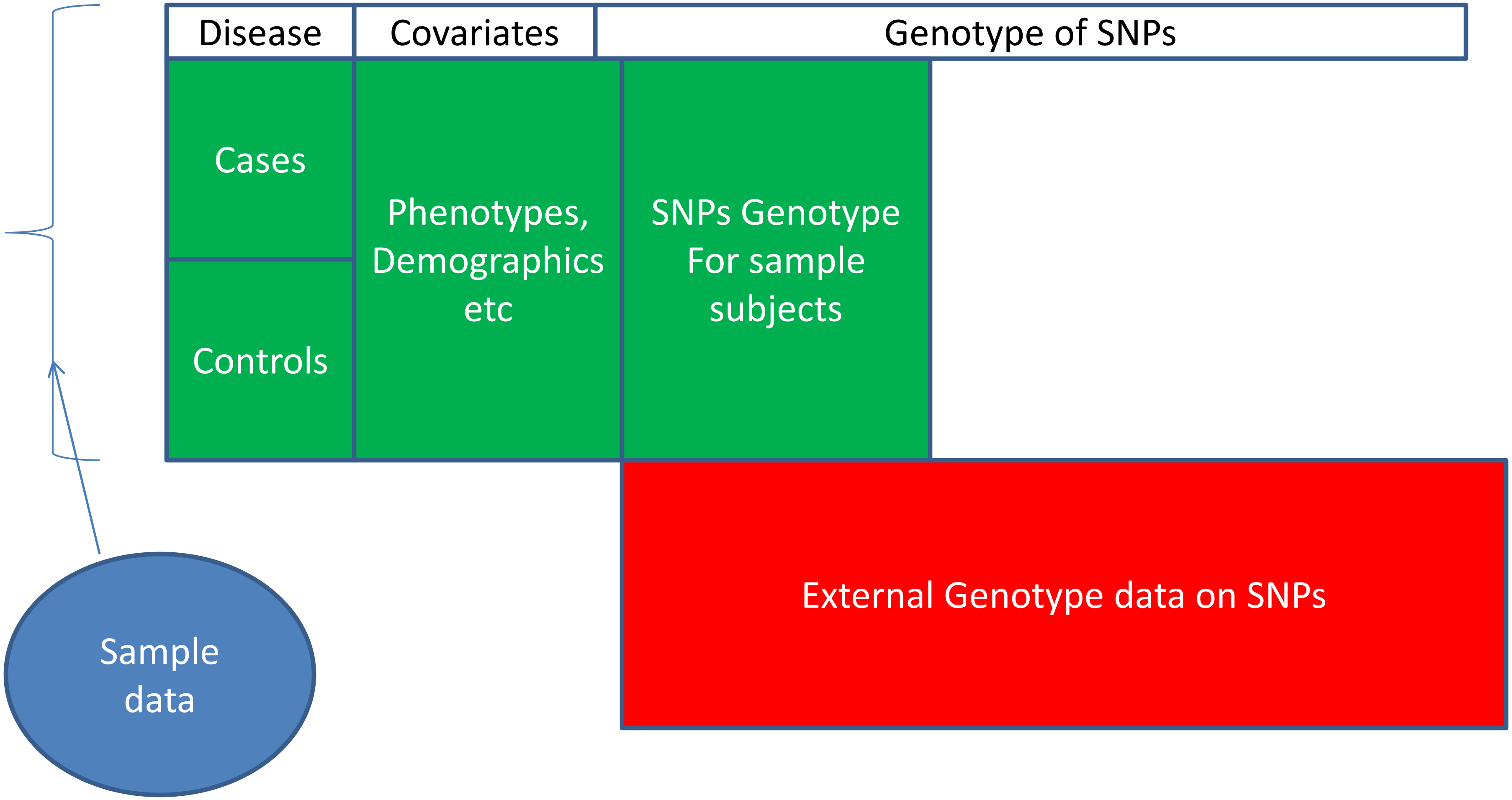
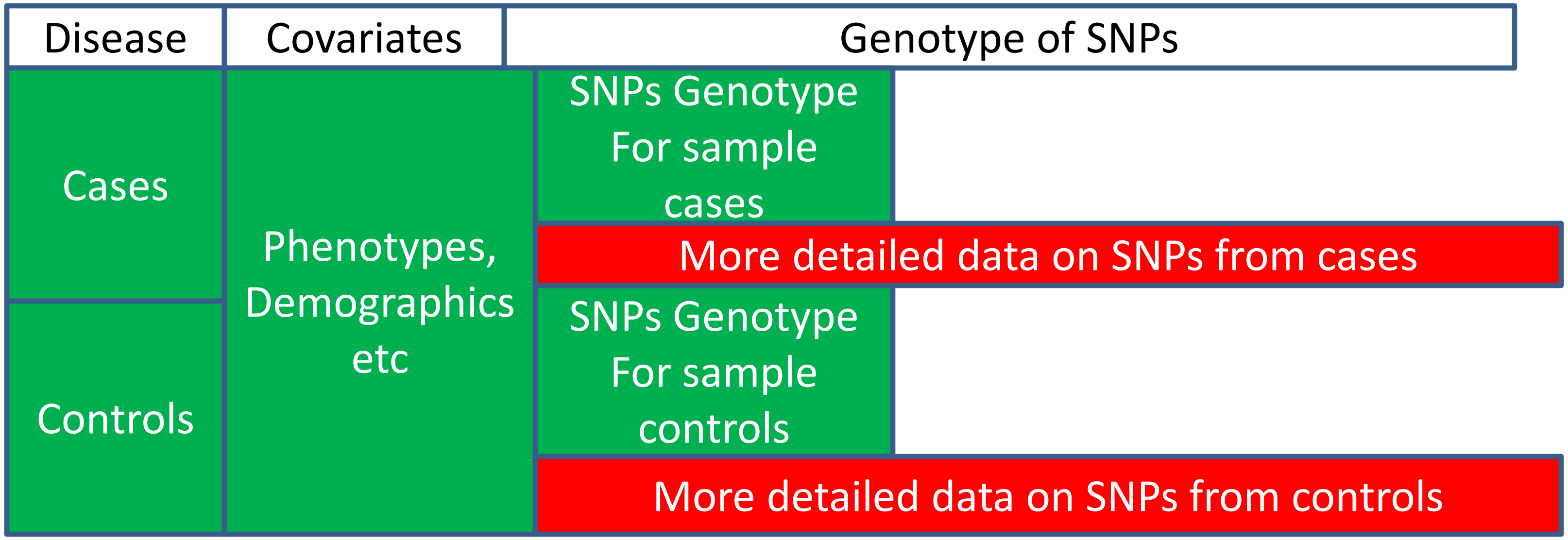

Disease
Covariates
Genotype of SNPs
Cases
Controls
Phenotypes, Demographics etc
SNPs Genotype For sample subjects
External Genotype data on SNPs
Sample data

Figure 2: An alternative self-contained scheme for GWAS studies



Disease

Covariates

Genotype of SNPs

Cases

Controls

Phenotypes, Demographics etc

SNPs Genotype For sample cases

More detailed data on SNPs from cases

SNPs Genotype For sample controls

More detailed data on SNPs from controls