Multiple Imputation as a Missing Data Machine

Jaap Brand[1,2], Stef van Buuren[2], Erik M. van Mulligen[1,3], Teun Timmers[1], Edzard Gelsema[1]
[1]Dept. of Medical Informatics, Erasmus University Rotterdam, The Netherlands
[2]Dept. of Statistics, TNO Prevention and Health, Leiden, The Netherlands
[3]University Hospital Dijkzigt, Rotterdam, The Netherlands

This paper deals with problems concerning missing data in clinical databases. After signalling some shortcomings of popular solutions to incomplete data problems, we outline the concepts behind multiple imputation. Multiple imputation is a statistically sound method for handling incomplete data. Application of multiple imputation requires a lot of work and not every user is able to do this. A transparent implementation of multiple imputation is necessary. Such an implementation is possible in the HERMES medical workstation. A remaining problem is to find proper imputations.

INTRODUCTION

The occurrence of missing data is a pervasive problem in clinical data analysis. Missing data can have many causes: respondents may be unwilling to fill in all items in a questionnaire, equipment can become defective, loss to follow up, and so on. Problems that are associated with incomplete data are: (1) cases with missing data may differ systematically from complete cases so that the sample is no longer representative. (2) less information is gathered than was intended, resulting in decreased power in statistical testing, and (3) many conventional statistical methods for complete data are not applicable anymore. Despite great effort that may have gone into collecting data, incomplete data are a fact of life.

In practice there are several methods to tackle the missing data problem. However, most of these methods have serious disadvantages. Three popular methods are:

(1) The deletion of incomplete cases. Simplicity is the main advantage of this method. However, an important disadvantage of this method is the potential loss of costly collected data. Moreover, estimators may be strongly biased, when incomplete cases differ systematically from complete cases.

(2) The development of adapted statistical methods for a postulated missing data mechanism. A theoretically elegant method is the Expected Maximalistion (EM) algorithm [1]. In this method, an explicitly defined missing data mechanism is combined with the selected sample model into a likelihood function. The parameters of the sample model are estimated with maximum likelihood. When the postulated missing data mechanism is correct, the results derived with EM are valid. However this method requires much statistical expertise and often specialised computer programs are required. Moreover EM is sometimes mathematically intractable.

(3) Completion filling in reasonable values for the missing data. An important advantage of this method is that after filling in the missing data, conventional methods for analyzing complete data can be applied. However, the disadvantages of this method are that it results in too small confidence intervals and correlations that are strongly biased, caused by the fact that the values filled in, are treated automatically as if they were known.

Clearly there is a need for an easy, generally applicable and statistically sound method. Such an method is multiple imputation as proposed by Rubin [2]. Multiple imputation is a very promising method and is the state of the art [3].

In this paper, we outline the general idea behind multiple imputation. Next we discuss the problem of implementing multiple imputation in a transparent way and propose how this problem can be solved by implementing multiple imputation in the HERMES medical workstation. Finally we discuss the remaining difficulties with multiple imputation, which still require research.

MULTIPLE IMPUTATION

The main goals of multiple imputation are taking into account the uncertainty about the missing data in a proper way and application of existing statistically methods for complete data. This can be done by filling in each missing value *m* times (*m*>=2), resulting in *m* completed data sets. When the fraction of missing information is modest, *m*=5 is sufficient [4]. The completed data sets are analyzed separately with the requested complete data method. Finally the *m* intermediate results are combined into one result. The flow of operations is illustrated in Figure 1.

0195-4210/94/$5.00 © 1994 AMIA, Inc.

The uncertainty about the missing data is reflected by the mutual variation between the imputed data sets. Little mutual variation between the imputed data sets means that there is little uncertainty about the missing data while much mutual variation between the imputed data sets means that there is much uncertainty about the missing data. Proper imputations can be obtained by drawing the values to impute Y^*_{mis} from the predictive distribution $P(Y_{mis} | Y_{obs}, R)$, R being the response indicator and Y_{mis} and Y_{obs} respectively the missing and observed part of the data Y.

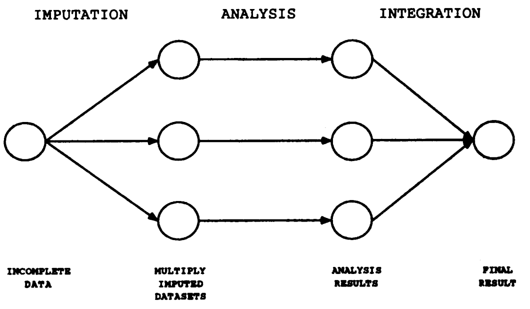

Figure 1. *Schematic representation of multiple imputation with m = 3*

The combination of the *m* intermediate results into one result can be performed by simple and sound procedures. The final parameter estimations are derived by averaging the intermediate parameter estimates. The uncertainty reflected by the variance between the imputations is taken into account by the final estimations of the variances and the p-values. The following three sources of uncertainty are taken into account: (1) the sample variation, (2) the missing data mechanism, and (3) the finite number of imputations used. The finite number of imputations is also a source of uncertainty, because from repeated application of the multiple imputation algorithm, different final results are obtained.

Application of multiple imputation requires much work for the user. For instance when a user applies multiple imputation to linear regression, he has to find *m* proper imputations for the incomplete data. The *m* completed data sets have to be analyzed separately with a statistical package like BMDP or SPSS. The *m* different intermediate results the user have to be combined into one result, with explicit formulas. Not every user is able to do this.

TRANSPARENT IMPLEMENTATION

To make multiple imputation applicable to a large group of users, multiple imputation has to be implemented in a transparent way, so that users can apply multiple imputation automatically. For a transparent implementation, the following is required:

- An imputator.
- Complete data analysis software (BMDP or SPSS).
- Filtering of the output, belonging to the complete data analysis.
- Pooling of the parameter estimators.
- An environment to integrate the modules.

Such an environment is the HERMES medical workstation [5]. The HERMES workstation is described below.

The design of the HERMES medical workstation accommodates the need to integrate different medical databases and software packages into a workstation. Its main goal is to offer the clinical user a friendly and transparent access to medical data and an easy use of existing software packages like BMDP, SPSS, WingZ and Harvard Graphics. The graphical user interface within HERMES has been developed with OSF/Motif and X11 and UIMX on a Hewlett Packard 9000/700 series workstation.

The architecture of the HERMES medical workstation is client-server based. Different applications in the HERMES environment can communicate with each other as client and server with a specially developed message language. A client, usually a graphical task-oriented user interface sends a request to a server which contains the functionality to solve the request. The results are sent back to the client. An application can act both as a client and as a server.

The communication between a client and a server is indirect, via a broker or a router. The broker uses a database to search the proper server that can handle the request. The broker database can be edited to add new servers. When alternative services have been defined in the database for a request, the broker can automatically select the most appropriate server.

The main advantages of the HERMES environment for the application programmer are: (1) Abstraction of the complexity of a program by division of the program into different modules. HERMES allows integration of these modules. (2) The possibility of using existing software, by encapsulating it with plugs into HERMES. Such a plug bridges between the

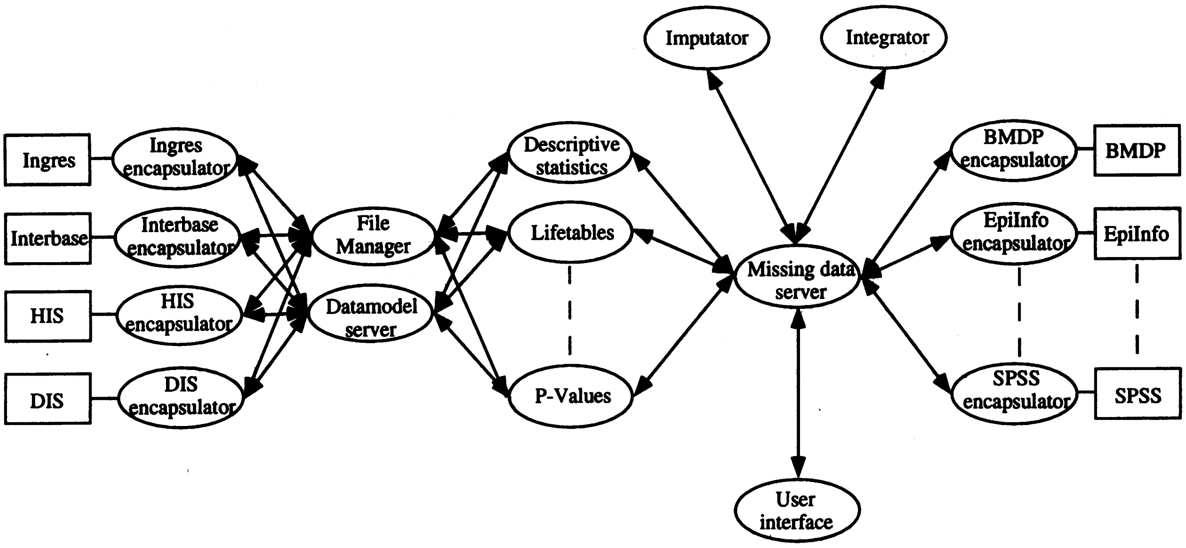

Figure 2. *Schematic representation of a client-server architecture for multiple imputation*

HERMES message language and the specific input and output formats of the application.This makes it possible to incorporate complete data analysis as required for transparent implementation of multiple imputation. How multiple imputation can be implemented transparently in HERMES is illustrated in Figure 2.

The missing data server receives the request from the statistical client and detects missing data. If no data are missing, the request is directly forwarded to the statistical service and the result is returned to statistical client. If data are missing, for each of the *m* completed data files generated by the imputator service the missing data service sends a request to the statistical service. Finally, all results are combined by the integrator service into one result and returned to the statistical client.

DISCUSSION

Multiple imputation is a promising method to solve the problems with missing data. It is possible to implement multiple imputation in a transparent way. The HERMES medical workstation is a suitable environment for this. A remaining difficulty is the derivation of the predictive distribution from which the imputations have to be drawn. The derivation takes the missing data mechanism and the sampling mechanism into account. In the literature there are three disjunct classes of missing data mechanisms.

There are two ways to take the sampling mechanism into account: model based and an data driven. In model based multiple imputation a statistical model (for instance the multivariate normal distribution), is used for the derivation of the predictive distribution. The concept of data driven imputation is to find an imputation that preserves the structure in the data as well as the uncertainty about this structure [6]. For instance, when in an incomplete data set two variables have an approximately quadratic relationship, the same variables should preserve this relationship after completing with a data driven imputation method.

The data driven imputation method has several advantages compared to model based imputation methods. Some advantages are:

- for different statistical methods for complete data, the same imputation method can be used.

- a data driven imputation method does not force the conclusions of subsequent analysis into a particular direction, a principle which is very important in statistics. If the data does not fit the assumed statistical model well, application of a model based multiple imputation method may lead to biased estimates.
- this method is very suitable for users with limited training in statistics.

The application of a data driven multiple imputation method is not always indicated. When there are good reasons to assume that the complete data can be described well by a statistical model, it is more suitable to apply a multiple imputation method based on this model, than applying a data driven imputation method. An intuitive reason for this is that using external knowledge such as a certain statistical model leads to more precise inferences. This assertion should be verified with simulation. Another reason is that by a simple model, model based imputation is much faster than data driven imputation.

Finally, a real data driven imputation algorithm is probably a utopia. Each imputation method always requires weak assumptions about the sampling mechanism. The idea of data driven imputation should be used on a gradual scale: an imputation method A is more data driven than another imputation method B when the sampling model used for method B is included in the sampling model used for method A.

To select the proper multiple imputation method and number of imputations *m*, especially for users with a limited training in statistics it is necessary to build in a selector in the missing data server, to come algorithmically to appropriate choices. The selector tests for each multiple imputations method a number of constraints and selects from the methods satisfying the constraints the most appropriate one. Constraints to be tested are for instance: The type of data, the number of variables, the percentage of missing data and some statistics concerning the sampling mechanism and missing data mechanism. Which constraints are to be tested should be investigated with simulation studies.

REFERENCES

[1] Little RJA, Rubin DB. Statistical analysis with missing data. Wiley, New York, 1987

[2] Rubin DB. Multiple imputation for nonresponse in surveys. Wiley, New York, 1987

[3] Tukey JW. Mixture modelling versus selection modelling with non-ignorable respons. In: Drawing Inferences from Self-Selected Samples. Springer-Verlag, New York, 1986

[4] Rubin DB. Multiple imputation in health - care data bases: an overview and some applications. Statistics in medicine. 1991;10:585-98

[5] Van Mulligen EM, Timmers T, van Bemmel JH. A new architecture for integration of heterogeneous software components. In: Methods of Information in Medicine, 1993:292-301

[6] Van Buuren S, Van Rijckevorsel JLA, Rubin DB. Multiple imputation by splines. In: Bulletin of the International Statistical Institute, Contributed Papers II, 1993:503-4