4

Handling missing data in risk factors in repeated measurements

Abstract

Public health researchers involved in data analysis are invariably confronted with non-response and resultant missing values. This missing information is a recurring problem in the statistical literature. In longitudinal and in panel studies, the problem of missing data is very common. To construct the risk career (e.g. smoking career) or to elicit the impact of the risk career on life expectancy free of cardiovascular disease and life expectancy with cardiovascular disease in a population, imputation of these missing values is essential. Without imputation, we may not be able to capture all events that we would like to relate to the risk factors at a nearby point in time. This chapter aims to propose a potential method to impute missing values of risk factors in repeated measurement studies. We have considered two risk factors-smoking status and systolic blood pressure (BP). Smoking status was selected to illustrate the application of the imputation method to categorical risk factors. BP was selected to illustrate the imputation method for continuous risk factors. We used the original Framingham Heart Study cohort to illustrate the proposed method. Over 48 years of follow-up, the overall percentage of respondents from whom no observations were missing in the Framingham Heart Study was 41.5. To impute the missing values on smoking, we followed the history of smoking of the relevant individual and replaced the missing values by the smoking status closest in time. The method we propose for imputing blood pressure values is based on a regression model, which is fitted to individual data and imputes mean values for each individual from each individual observation. The algorithm of the proposed method is easy to manipulate and to understand. We have validated our method and found that the proposed method is justifiable for long-period longitudinal data. Applying this method to impute missing values of other risk factors is possible. The imputation may result in better estimates of the required parameters than when the missing cases are omitted.

4.1 Introduction

Incomplete or missing data is a recurring problem in data analysis. In particular, public health researchers involved in data analysis are invariably confronted with non-response and resultant missing values (Zhou et al., 2001). Missing information may exist in any empirical study. In longitudinal or panel studies, missing data are very common (Molenberghs, 2001). The main reason is that in longitudinal studies, subjects may drop out early or be unavailable during one or more data collection episodes. During follow up, respondents may fail to return in longitudinal studies because of death, illness or disability. Alternatively, subjects may be unwilling or may fail to complete a section of the questionnaire due to lack of time or interest. If such attrition is random with respect to the dependant variables, it can be ignored from the analysis. Otherwise, the probability of dropout depends upon covariates and/or previously observed response, or on current unobserved responses (Little, 1995). However, though the causes and problems with missing values in longitudinal research have been identified (Little, 1995; Zhou et al., 2001; Molenberghs, 2001; West and Dawson, 2002; Wei and Shih, 2001) the methods available for imputation of missing values are quite limited.

Repeated measurements are quite common in clinical trials. However, the number of longitudinal studies is increasing in public health research to measure the health status of the elderly and the life course of non-communicable diseases (e.g. cardiovascular disease). In developed countries, there are many longitudinal studies, where multiple waves of measurements (panel) are available. The multiple waves of measurement bring multiple opportunities for missing data. Dropout and intermittent non-response are found in most longitudinal studies (Little, 1995). Because of the missing information, we may not be able to analyze time-dependent covariates (e.g. risk factors). Most researchers use a part of the follow-up data or only the baseline information. For example, to calculate the 'cardiovascular disease risk profile', Anderson et al., (1991) used the FHS follow-up period from 1968 to 1975 with at least one observation of the listed risk factors. Stamler et al., (1999) used baseline information to measure the impact of several risk factors on the incidence of cardiovascular disease. Although the disease incidence and mortality is recorded in exact time, during the follow-up the likelihood of missing risk factors is quite common. Therefore, analysis of the risk career of a disease process or mortality is difficult without imputation of missing values.

To construct the risk career (e.g. smoking career) or to see the impact of the risk career on life expectancy free of cardiovascular disease and life expectancy with cardiovascular disease, imputation of missing values can be useful. Without imputation, we may not be able to capture all events that we would like to relate with the risk factors at a nearby point in time. Hence, the objective of this chapter

is to describe a potential method to impute missing values of risk factors in a repeated measurement study. To this end, we considered two risk factors- *smoking status* and *systolic blood pressure* (BP). Smoking status was selected to illustrate the imputation method for categorical risk factors. BP was selected to demonstrate the imputation method for continuous risk factors. We used the original Framingham Heart Study (FHS) cohort to apply the proposed method. In this study, the same cohort of 5,209 subjects has been followed for 48 consecutive years. Several risk factors of cardiovascular disease, cardiovascular disease incidence and mortality were recorded during the follow-up period. After 48 years of follow up, 77 percent of the respondents who had participated in the first exam during 1948-51 had died. In total 15,800 (22 percent) exams or responses are missing from the potential 87,578 responses.

In the FHS, all events (cardiovascular disease and death) are recorded at the exact time. Three types of missing information could be identified. First, there were the respondents who did not participate in a particular exam (*missing type I).* The second type of missing data concerned risk factor information that failed to be recorded for some of the respondents even though the respondents were present in that exam (*missing type II*). The third type referred to the exams where risk factor information had been omitted completely for all subjects although the respondents had been present at the exam in question *(missing type III)*. For example, smoking status failed to be recorded at exam 6.

A common method for handling missing values is imputation. There are no universally applicable methods for handling missing data (Shih, 2002). Several methods of imputation are available, such as the hot deck approach, the regression method, multiple imputation and so on. Each method has its own advantages and disadvantages. The selection of the appropriate method depends on the assumptions made about the missing data mechanism. Many researchers use these methods for different data sets. Imputation methods for repeated measurement data are limited. This chapter describes the missing information in the FHS and subsequently illustrates the method proposed by us for imputing missing values.

We propose two methods for imputation: one for discrete variables and another one for continuous variables. In the case of a discrete variable, the history of the risk factor in question is followed after which the missing values are replaced by the risk factor status that is close in time. Continuous variables are imputed by fitting a regression model to each individual, which model is used to impute the missing value. That is, missing values are imputed based on the individually observed values. For an application, see Chapter 5.

Section 4.2 of this chapter describes the missing data mechanism. Section 4.3 provides an overview of missing data in the FHS. In this section, we have illustrated missing information on two risk factors: smoking status and systolic blood pressure. Section 4.4 describes the imputation methods that are proposed and applied to the

Framingham Heart Study. In subsection 4.4.1, we describe the imputation method applied for imputation of smoking status. The method we use for the imputation of BP is discussed in subsection 4.4.2. In Section 4.5, we validate our proposed methods. Section 4.6 concludes the chapter.

4.2 The missing data mechanism

The missing data mechanism is a concept often discussed when missing data occur. To understand the consequences of missing values and potential solutions for statistical analysis with missing values, some idea of why and how missing values occur is needed. The missing data mechanism is described in two parts: (i) substantive issues and (ii) problems. In the first part, we describe the types of missing data mechanisms. The second part deals with associated problems with missing data.

4.2.1 Substantive issues

In modern statistical exercise, the occurrence of missing values is usually viewed as random phenomenon (Schafer, 2001). Little and Rubin (1987) define the occurrence of missing values by three unique types of missing data mechanisms. Missing values of a random variable Y can be (i) *Missing completely at random* (MCAR), (ii) *Missing at random* (MAR) and (iii) *Nonignorable*. We assume that the y_1 portion of y (a realization of Y) is missing and the rest y_2 (y-y_1) is observed.

(i) *Missing completely at random* (MCAR)
Missing data is considered to be MCAR when, given two variables, say, X and Y, the probability of response is independent of variables X and Y. That is, cases with complete data are indistinguishable from cases with incomplete data. MCAR is the process in which the probability of data being missing is independent of both observed measurements (e.g. baseline covariates, observed responses) and unobserved measurements (those that would have been observed if the respondent had remained in the study). Under an MCAR process, whether or not a variable is observed does not affect its distribution (Little et al., 2000). This can be written as-

$$Pr(Y=y \mid y_1 \text{ missing}, y_2 \text{ observed}) = Pr(Y=y \mid y_2 \text{ observed}), \text{ where } y_1 \text{ and } y_2 \in Y$$

As an example, suppose serum cholesterol level (SCL) and age are variables of interest for our study. If the likelihood that a respondent would provide his or her serum cholesterol level is the same for all individuals regardless of their SCL or age, then the missing data is considered to be MCAR. Under MCAR, the observed

responses form a random sub-sample of the sampled responses (Rubin, 1976). Therefore, when data are MCAR there is no impact on bias and most standard approaches of analysis are valid.

(ii) *Missing at random* (MAR)

Missing data is considered to be MAR when, given two variables, X and Y, the probability of response depends on X but not on Y. That is, the cases with incomplete data differ from cases with complete data, but the pattern of data missingness is traceable or predictable from other variables in the database rather than being due to the specific variable on which the data is missing. Under MAR, the probability of missing depends on the observed data rather than on the unobserved data. The observed responses are a random sub-sample of the sampled values within a subclass defined by the observed data. Therefore, MAR is a more relaxed condition, assuming only that missing and observed distributions of Y are identical, conditional on predictor X, i.e.,

$$Pr(Y=y \mid y_1 \text{ missing}, y_2 \text{ observed}, x \text{ observed}) = Pr(Y=y \mid y_2 \text{ observed}, x \text{ observed})$$

In this situation, the missing data mechanism depends only on the covariates X and is classified as covariate-dependent missing (William, 2000). For example, again using the example of serum cholesterol level and age, if the likelihood that an individual would provide his or her SCL varied according to an individual's age, the missing data is considered to be MAR. Most of the missing data methods are designed under this assumption.

(iii) *Nonignorable*

When, given two variables, X and Y, the probability of response depends on X and possibly Y, missing data is considered to be nonignorable i.e.

$$Pr(Y=y \mid y_1 \text{ missing}, y_2 \text{ observed}, x \text{ observed}) = Pr(Y=y \mid y_1 \text{ observed}, y_2 \text{ missing}, x \text{ observed})$$

In other words, if the missing mechanism is neither MCAR nor MAR, then it is nonignorable. Under nonignorable, the pattern of data missingness is non-random and is not predictable from other variables in the database. An example of this would be if the likelihood of a respondent providing his or her SCL varies according to a person's SCL (observed and missing) and age (covariate).

In repeated measurement data, there are ultimately two types of missing information: bounded missing and missing due to dropout. Bounded missing is defined as a missing value that has at least one observed value before, and at least one observed value after the period in which it is missing (SOLAS, 1999). Missing

due to dropout refers to the case where a participant is dropped from the follow-up and never comes back in that survey. The following table shows an example of bounded missing data and data missing due to dropout. The variables Exam1 to Exam5 are a set of longitudinal measures of SCL for 5 respondents. The SCL of the first respondent is missing at exam 2 and exam 3, which is bounded missing. Similarly, the second respondent's SCL is bounded missing at exam 3. Respondent 4 has dropped out from exam 3 onward, which is missing due to dropout.

Table 4.1 An example of repeated measurement data (a hypothetical example of cholesterol level)

Respondent Identification Number	Exam1	Exam2	Exam3	Exam4	Exam5
1	212	*	*	223	240
2	190	185	*	232	215
3	240	212	227	218	235
4	222	231	-	-	-
5	185	198	175	192	200

'*' bounded missing
' -' missing due to dropout

4.2.2 Problems

The issue of missing data is the subject of increasing debate in contemporary statistics. Hence missing data is a problem in almost all areas of empirical research (Rubin, 1996). In any given study, missing data can have many causes. For instance, respondents may be unwilling to answer some questions (which is called item non-response) or may refuse to participate in a study (which is called unit non-responses) and so on. In a longitudinal study, there is always a higher chance of missing observations during follow–up than in a cross-sectional study. Thus, many longitudinal studies suffer from attrition; that is, from subjects dropping out prematurely (Laird, 1988; Hedeker and Gibbons, 1997; William, 2000).

Often, the question arises of why the cases with missing values should not simply be deleted rather than imputing values at all. Little and Rubin (1987), among other researchers, have demonstrated the dangers of simply deleting cases. Basically, case deletion strategies assume that the deleted cases are a relatively small proportion of the entire data set and are representative of it—that is, cases will then be missing completely at random. In most research settings, however, missing data are indicative of some pattern and cannot safely be assumed to be at random. In such circumstances, deletion can introduce substantial bias into the study (Laird, 1988). Moreover, the loss in sample size can appreciably diminish the statistical

power of the analysis. Missing values lead to less efficient estimates because of the reduced size of the database, while standard complete-data methods of analysis no longer apply. When cases are deleted from the data because of one or more variables with missing values, the number of remaining cases may be small even if the proportion of missing data is small for each variable.

It is obvious that if some observations are incomplete or missing, there is more uncertainty in inference than if the data had been complete. The main reasons are the smaller sample size, and the bias introduced when the pattern of missing data results from some process. Hence, analysis of data sets with incomplete information is more problematic than analysis of complete data sets. Therefore, in the presence of non-response that cannot plausibly be considered to be completely random, estimates of population parameters are subject to potential bias (U.S. Census Bureau, 2002; Shih, 2002).

4.3 Missing data in the FHS

We have used the original FHS cohort to illustrate the proposed method. The FHS consisted of 5,209 respondents (45% male) from a sample of adults aged 28 to 62 years residing in Framingham, Massachusetts between 1948 and 1951. The FHS is a long time longitudinal study. We have used the first 48 years of follow-up, from exam 1 in 1948 to exam 24 in 1998 of the FHS. For this type of study, missing information in any wave is possible because of (i) bounded missing, (ii) loss to follow-up if for other reasons than death or (iii) death. In the FHS, death is not a reason for loss of follow-up. We assumed that missing observations were only possible before death. Subsection 4.3.1 examines the total missing cases and 4.3.2 illustrates the missing values of two risk factors in the FHS.

4.3.1 Total missing cases in the FHS

Table 4.2 shows the distribution of absent or missing cases according to the total number of exams missing until end of follow-up or death. The first column of this table represents the number of exams missing per individual and column 2 is the frequency of missing cases. Out of a total of 5,209 respondents, 902 (17.32 percent) had missed only 1 exam. For 30 persons (0.58 %), 23 out of 24 exams were missing. A total of 2,160 (41.47 percent) persons participated in all exams before death or throughout 48 years of follow-up.

Table 4.2 Total number of exams missing

Total number of absent exams	Frequency	Percent	Total number of absent exams	Frequency	Percent
0	2160	41.47	13	48	0.92
1	902	17.32	14	37	0.71
2	441	8.47	15	34	0.65
3	321	6.16	16	38	0.73
4	223	4.28	17	35	0.67
5	163	3.13	18	21	0.40
6	140	2.69	19	22	0.42
7	100	1.92	20	15	0.29
8	118	2.27	21	22	0.42
9	90	1.73	22	21	0.40
10	89	1.71	23	30	0.58
11	71	1.36			
12	68	1.31	Total	5209	100

Table 4.3 shows the number of missing cases for each exam due to non-participation in the exam. The 1st column is the order of exam; the 2nd column shows the cumulative number of deaths; the 3rd column indicates the number of survivors; the 4th column is the number of persons who participated in the exam; the 5th column is the number of respondents who did not participate in the exam and the 6th column is the percentage of missing cases for reasons other than death.

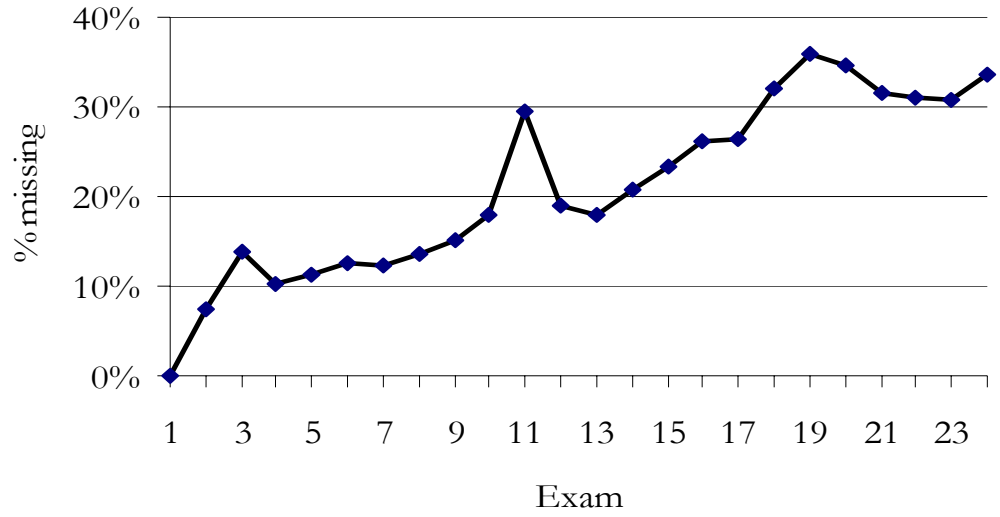
A total of 5,209 respondents attended the first exam in the FHS. The death of a participant was recorded in the next most recent exam. If a subject died between x^{th} and $(x+1)^{th}$ exam date, his or her death would be recorded at $(x+1)^{th}$ exam. For example, of the 5,209 persons who participated in the first exam, 33 died before the 2nd exam and 384 did not participate in the 2nd exam for reasons other than death. The remaining 4,792 (i.e. 5,209-33-384) individuals participated in the second exam (Table 4.3). Out of a total of 87,578 potential observations, some 71,778 observations were recorded. The non-response rate in the Framingham Heart Study was 22 percent. Table 4.3 and Figure 4.1 show that, as the years go by, the percentage of missing responses increases. Almost 7 percent of the respondents were missing at the 2nd round interview and about 34 percent were missing from the 24th exam. A sudden increase of the number of non- participants could be observed at exam 11. The NHLBI (National Heart, Lung, and Blood Institute) had directly staffed and funded the Framingham Study original cohort for 20 years. At exam 11, direct funding was running out, and the NHLBI was going to close the examinations. They were therefore unable to complete the full number of exams. The study ran on a variety of foundation type supports for a few years, and was

then picked up again and funded under a NHLBI contract with Boston University (Sorlie, personal communication). After 48 years of follow-up, 1250 respondents (almost 24 percent) of the initial members, are still under observation. A participant may be absent from one or several exams, and although the exact date of an event (if any occurred) will be recorded in any of the subsequent exams, the risk factor values (e.g. blood pressure) in the missing exam(s) will remain missing.

Table 4.3 Missing cases in given exam

Exam	Cumulative number of deaths	Number of survivors	Number of cases participated	No participation	Missing (%)
1	-	5209	5209	0	0
2	33	5176	4792	384	7
3	87	5122	4416	706	14
4	151	5058	4541	517	10
5	230	4979	4421	558	11
6	332	4877	4259	618	13
7	429	4780	4191	589	12
8	548	4661	4030	631	14
9	688	4521	3833	688	15
10	833	4376	3595	781	18
11	1017	4192	2955	1237	30
12	1186	4023	3261	762	19
13	1395	3814	3133	681	18
14	1583	3626	2871	755	21
15	1776	3433	2632	801	23
16	2024	3185	2351	834	26
17	2244	2965	2179	786	27
18	2524	2685	1825	860	32
19	2808	2401	1541	860	36
20	3063	2146	1401	745	35
21	3283	1926	1319	607	32
22	3517	1692	1166	526	31
23	3728	1481	1026	455	31
24	3959	1250	831	419	34
Total			71,778	15,800	22

Figure 4.1 Missing cases (%) in given exam



4.3.2 Missing risk factors

In the FHS, the exact timing of events is known. The risk factors are missing in some exams. A risk factor variable can be binary (e.g. smoking status, yes/no) or continuous response (e.g. blood pressure). To illustrate the missing statistics in the FHS we considered two selected risk factors: *smoking* and *blood pressure*. It is possible to visualize the missing data in the FHS for other risk factors in the same manner.

In this section, the patterns of missing information on smoking and blood pressures in the FHS are described briefly. As we discussed earlier (Section 4.1), there are three types of missing cases in the FHS. Missing type I occurs when some respondents do not participate in that exam. Missing type II occurs when the risk factor information is not collected although the respondent participates in the exam. Missing type III occurs when the risk factor information for all respondents is not conducted in the relevant exam.

Smoking

Smoking status of the FHS respondents was recorded by asking two questions: currently smoking (yes or no) and number of cigarettes per day (24 hours). We have demonstrated current smoking status only. The smoking status recorded at each exam is given in Table 4.4. There is no information on smoking status at exam numbers 2, 3, 6 and 16. After 48 years of follow-up, some 71,778 recorded responses had been collected, 48.54 percent of which indicated that the respondent smoked,, 29.89 percent were recorded as non-smokers and the remaining 21.57 (type I 19.15 and type II 2.42 percent) percent were missing. Of the potential observations, 20.96 percent were missing type III i.e. smoking status failed to be

recorded at exams 1, 2, 3 and 6. At the time of entry or at the first exam, 57.3 percent of the respondents were current smokers and only 0.6 percent was missing (missing type I). At the 24th exam, after 48 years of follow-up only 4.16 percent replied that they were current smokers and 62.24 percent indicated that they were non-smokers. Smoking status was missing for 33.52 percent respondents in last exam.

Table 4.4 Smoking status by exam, FHS

Exam	Non-smoker (%)	Smoker (%)	Missing (%)			Total survivors
			Type I	Type II	Type III	
1	42.02	57.34	0.00	0.63	-	5209
2	-	-	-	-	100.00	5176
3	-	-	-	-	100.00	5122
4	41.89	46.58	10.22	1.30	-	5058
5	41.65	46.98	11.21	0.16	-	4979
6	-	-	-	-	100.00	4877
7	43.66	43.89	12.32	0.13	-	4780
8	44.65	41.82	13.54	0.00	-	4661
9	46.07	38.71	15.22	0.00	-	4521
10	48.01	34.12	17.85	0.00	-	4376
11	26.00	13.17	29.51	31.32	-	4192
12	52.97	28.09	18.94	0.00	-	4023
13	55.45	26.66	17.86	0.03	-	3814
14	55.93	23.14	20.82	0.00	-	3626
15	56.25	20.33	23.33	0.09	-	3433
16	-	-	-	-	100.00	3185
17	51.20	15.38	26.51	6.91	-	2965
18	56.46	11.21	32.03	0.30	-	2685
19	55.10	8.79	35.82	0.29	-	2401
20	56.34	8.29	34.72	0.65	-	2146
21	62.25	6.23	31.52	0.00	-	1926
22	63.30	5.50	31.09	0.12	-	1692
23	64.69	4.59	30.72	0.00	-	1481
24	62.24	4.16	33.52	0.08	-	1250
Total	48.54	29.89	19.15	2.42	20.96	87,578

Note: Missing type I: no participation in the exam
Missing type II: participated in the exam but missing in reporting smoking status
Missing type III: smoking status is not recorded

Systolic blood pressure (BP)

We analyzed the BP in mm Hg recorded by the second examiner. The systolic blood pressure status at each exam is given in Table 4.5. There are no type III missing responses in the BP recording of this study, only type I and type II. BP had been recorded for 92.2% respondents at the first exam; the remaining 7.83% were missing, i.e. no BP measurement was taken. Overall, 22.93 percent (type I 18.04 percent and type II 4.89 percent) values of BP were missing during the follow-up period, even though the respondents in question had been examined.

Table 4.5 Systolic blood pressure by exam, FHS

Exam	SPS (%)	Missing (%)		Total survivors
		Missing-type-I	Missing-type-II	
1	92.17	0.00	7.83	5209
2	87.60	7.42	4.98	5176
3	80.38	13.78	5.84	5122
4	78.65	10.22	11.13	5058
5	79.43	11.21	9.36	4979
6	83.92	12.67	3.40	4877
7	85.98	12.32	1.69	4780
8	84.47	13.54	2.00	4661
9	83.43	15.22	1.35	4521
10	81.86	17.85	0.30	4376
11	38.50	29.51	0.32	4192
12	80.86	18.94	0.20	4023
13	81.65	18.78	0.50	3814
14	78.19	20.82	0.99	3626
15	75.12	23.33	1.54	3433
16	72.72	26.19	1.10	3185
17	70.46	26.51	3.04	2965
18	67.37	32.03	0.60	2685
19	63.64	35.82	0.54	2401
20	65.14	34.72	0.14	2146
21	68.28	31.52	0.21	1926
22	69.00	31.09	0.00	1692
23	69.07	30.72	0.20	1481
24	46.24	33.52	20.24	1250
Total	77.07	18.04	4.89	87,578

4.4 Missing value imputation

Imputation is the name given to any method whereby missing values in a data set are filled-in with reasonable estimates (SOLAS, 1999). The ultimate goal of any imputation method is to produce a complete data set. That complete data set can be analyzed using complete-data inferential methods. There are two common procedures for dealing with missing data. Either all data records in which any variable value is missing are excluded or 'plausible' values are substituted for the missing items (Goldstein and Woodhouse, 1996). Methods based upon the use of plausible or imputed values underlie most recommended procedures, and detailed discussions are given by Rubin (1987), Little and Rubin (1987) and Little (1993). To analyze the risk factor career in Framingham Heart Study, we need imputation. In the literature, there are several popular methods for handling missing data. Much of the literature involving missing data in public health research pertains to the various methods developed to handle the problem. For details of the available methods to impute missing values we refer to Rubin (1987); Hedeker and Gibbons (1997); Heitjan (1997); Little (1993); Little and Schenker (1995); Little and Rubin (1987). However, there are no efficient methods for imputing missing data in repeated measurement.

Handling missing values for repeated measurement, public health researchers frequently use the *last value carried forward* (LVCF) method. The last observed value is used to fill in missing values at a later point in the study. LVCF makes the assumption that the response remains constant at the last observed value. In any intervention study (e.g. clinical trial), this can be biased if the timing and rate of withdrawal is related to the treatment (William, 2000). For example, in the case of degenerative diseases, using the last observed value to impute missing data at a later point in the study can produce biased results.

Multiple imputation[1] has good statistical properties, although this method has not yet been extensively used in longitudinal data analysis. Moreover, the method had not been used heretofore with repeated measurement data, especially follow-up data over a long period. The main reason for this is that missing values in repeated

[1] Multiple imputation: MI is a Monte Carlo technique in which the missing values are replaced by m>1 simulated versions, where m is typically small (e.g. 3-10). In Rubin's (1987) method for 'repeated imputation' inference, each of the simulated complete data sets is analyzed by standard methods, and the results are pooled to produce estimates and confidence intervals that incorporate missing-data uncertainty. Primarily, Rubin (1987) addresses potential uses of MI for large public-use data files from sample surveys and censuses. With the advent of new computational methods and software for creating MI's, however, the technique has become increasingly attractive for researchers in different backgrounds whose investigations are hindered by missing data. Schafer (1997) documents these methods in a recent book on incomplete multivariate data.

measurement data are mostly correlated with covariates that make the explanation of the missing mechanism critical. The lengthy procedure of multiple imputation and covariates dependency in repeated measurement data are an impediment to the application of multiple imputation. Little (1995) modeled the dropout mechanism in a repeated measurement study. He used a pattern mixture model for various missing data mechanisms. Wei and Shih (2001) proposed a partial imputation approach to analyze the repeated measurements with dependent dropouts. Their method is an extension of LVCF. Before choosing a missing data handling approach, we might keep in mind that one of the desired outcomes is maintaining (or approximating as closely as possible) the shape of the original distribution of responses. Some incomplete data handling methods do a better job of maintaining the distributional shape than others.

Because of the limited use and disadvantages of available methods in longitudinal data, we have proposed two approaches: one for categorical risk factors and another for continuous risk factors. For categorical risk factors, we propose an algorithm which is an extension of LVCF (say, ELVCF). For continuous risk factors, we propose a regression method which is an extension of the regression method (say, ERM). These methods are applicable for categorical and continuous missing data in repeated measurement study.

We assumed that the missing data in the FHS were MAR and that the pattern was monotone[2]. We have illustrated the method with an example. To describe the ELVCF method, we treated smoking as a categorical variable: smokers and non-smokers. Based on several assumptions, we imputed values for missing smoking status. For the ERM method, we considered systolic blood pressure to be a continuous variable. We imputed the missing values of BP based on the observed distribution of BP. The imputation procedures are illustrated with an example.

4.4.1 Smoking status imputation

We first treated smoking as a categorical variable: smokers and non-smokers. People were viewed as having a 'smoking career' throughout the observation period. They could drop in and out of smoking status. There was no 'ex-smoker' category. Now, not everyone turned up at all rounds. We imputed likely values, and flagged these according to the following algorithm, assuming a missing value at exam or round *i*. An individual could be absent from one or from several consecutive exams or drop out. We therefore limited our method to imputing missing values for a maximum of two consecutive missing exams.

[2] A monotone missing data pattern occurs when the variables can be ordered, from left to right, such that a variable to the left is at least as observed as all variables to the right.

Box Smoking code after and before imputation

A note on coding:

1: smoking at that exam (from original data)
0: not smoking at that exam (from original data)
2=smoking, imputed
3=non-smoking, imputed
9=missing
2: conditions are as follows:

- **a.** If i^{th} round missing (9) and $(i\text{-}1)^{th}$ round smoked (1) and $(i+1)^{th}$ round smoked (1) then **code=2**
- **b.** If i^{th} round missing (9) and $(i\text{-}1)^{th}$ round smoked (1) and $(i+1)^{th}$ round not smoked (0) then **code=2**
- **c.** If i^{th} round missing (9) and $(i\text{-}1)^{th}$ round smoked (1) and $(i+1)^{th}$ round missing (9) then **code=2**

3: conditions are as follows:

- a. If i^{th} round missing (9) and $(i\text{-}1)^{th}$ round not smoked (0) and $(i+1)^{th}$ round smoked (1) then **code=3**
- b. I i^{th} round missing (9) and $(i\text{-}1)^{th}$ round not smoked (0) and $(i+1)^{th}$ round not smoked (0) then **code=3**
- c. If i^{th} round missing (9) and $(i\text{-}1)^{th}$ round not smoked (0) and $(i+1)^{th}$ round missing (9) then **code=3**

9: conditions are as follows:

- a. If i^{th} round missing (9) and $(i\text{-}1)^{th}$ round missing (9) and $(i+1)^{th}$ round missing (9) then **code=9**

Following the description and code scheme shown in the Box, the algorithm for imputation of missing smoking status can be described as follows:

i. If we have consistent observation at round *i*-1 and at round i+1, smoking status is imputed as observed status at both observed rounds.
ii. If we have inconsistent information at round *i*-1 or at round *i*+1, smoking status is imputed based on the assumption that the state changes at midpoint.
iii. If we have one missing round, the status of the previous round is imputed (i.e. carrying forward)
iv. If we have two consecutive missing rounds, the status of the round nearest in time is imputed (i.e. carrying forward for the first and carrying backward for the second)
v. If three or more consecutive rounds are missing, only the middle record is assigned the status 'missing'.
vi. If the information we have is all 'right censored' due to the fact that the subject failed to show up at all subsequent rounds, i.e. up to and including l the 24^{th} exam, the observations are recorded as missing.

Table 4.6 Smoking status of an individual

Exam	ID	Live	Age	Smoking code before imputation	Smoking code after imputation
1	16	1	33	0	0
2	16	1	36	9	3
3	16	1	38	9	2
4	16	1	39	1	1
5	16	1	42	1	1
6	16	1	43	9	2
7	16	1	45	1	1
8	16	1	47	1	1
9	16	1	49	1	1
10	16	1	51	1	1
11	16	1	53	0	0
12	16	1	56	1	1
13	16	1	59	1	1
14	16	1	.	9	2
15	16	1	.	9	9
16	16	1	.	9	9
17	16	1	.	9	9
18	16	1	.	9	9
19	16	0	.	.	.
20	16	0	.	.	.
21	16	0	.	.	.
22	16	0	.	.	.
23	16	0	.	.	.
24	16	0	.	.	.

Note: in column under the name age ' .' is missing type I and other columns it is system missing as the individual died

The smoking history of an individual (case=16) participant in the FHS is shown in Table 4.6. This individual entered the FHS at age 33 as a non-smoker, was absent from the 2nd and 3rd exam (smoking status was not gathered at the 2nd and 3rd exam), and was a smoker at the 4th and 5th round. He was missing at the 6th round, a smoker at exam 7 to exam10, a non-smoker at the 11th exam, but a smoker again at the 12th and 13th exam; from the 14th exam onward he was absent from the survey and his death (DTH) was recorded in 19th exam, i.e. he died between the 18th and 19th exam. According to our algorithm and coding scheme (Box), the missing values at the 2nd and 3rd exam were imputed as 3 and 2 respectively. He was absent from the 6th exam, which value is replaced by 2. As he was absent after the 13th

exam onward, all missing values remained missing after the 14th exam to the time of his death (i.e. before the 19th exam).

In the FHS, 87,578 smoking status values have been collected in all potential exams (from exam 1 to exam 24). Following the algorithm described above, we imputed the missing values of the smoking variables in the FHS. An explanation of the coding used for smoking status before and after imputation is given in the Box.

Table 4.7 shows the smoking status distribution of males and females of the FHS cohort before and after imputation. A total of 33,289 (38.0 percent) observations of smoking status were missing in the FHS. Of these 11,905 (13.6 percent) responses of non-smokers and 10,438 (11.9 percent) responses as smokers were able to be imputed. After imputation, 10,946 (12.5 percent) observations were still missing. Following the exclusion of the missing observations, before imputation 51.7 percent of males and 28.4 percent of females were shown to be smokers. After imputation, 55.2 percent of males and 30.1 percent of females were found to be smokers.

Table 4.7 Smoking status before and after imputation, FHS

Smoking status	Male		Female		Total	
	% of total values (N)	% of non-missing values	% of total values (N)	% of non-missing values	% of total values (N)	% of non-missing values (N)
Before imputation						
Non-smoker	30.2 (10949)	48.3	44.1 (22653)	71.6	38.4 (33602)	61.9
Smoker	32.3 (11705)	51.7	17.5 (8982)	28.4	23.6 (20687)	38.1
Missing	37.4 (13556)	-	38.4 (19733)	-	38.0 (33289)	-
After imputation						
Non-smoker						
Recorded	30.2 (10949)	34.1	44.1 (22653)	50.9	38.4 (33602)	43.8
Imputed	9.5 (3438)	10.7	16.5 (8467)	19.0	13.6 (11905)	15.5
Total	39.7 (14387)	44.8	60.6 (31120)	69.9	52.0 (15265)	59.4
Smoking						
Recorded	32.3(11705)	36.5	17.5 (8982)	20.2	23.6 (20687)	27.0
Imputed	16.6 (6003)	18.7	8.6 (4435)	10.0	11.9 (10438)	13.6
Total	48.9(17708)	55.2	26.1 (13417)	30.1	35.5 (31125)	40.6
Missing	11.4 (4115)	-	13.3 (6831)	-	12.5 (10946)	-

Note: N is the number of responses

4.4.2 Systolic blood pressure imputation

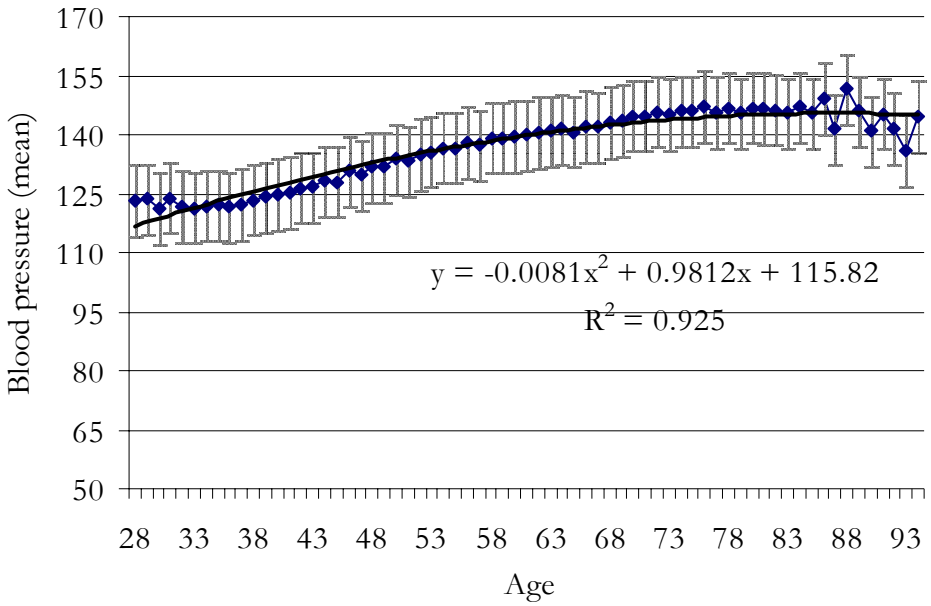
Blood pressure is measured as a continuous variable in the FHS. Before imputation, we plotted the mean value of BP by age of respondents, because BP is age dependent. We were interested to see whether individual prediction equations could

be derived from which missing values could be imputed. We explored the shape of the observed mean of BP against age. We found that the shape fit a quadratic relationship of BP as age changed, i.e. non-linear regression, where BP was a response variable and age and age*age were independent variables. The observed means of BP by age are shown in Figure 4.2. For the observed plot, the *X*- axis represents respondent age at the time of exam and *Y*-axis represents BP mean by age at that corresponding exam. After plotting the observed BP mean against age, the trend line drawn followed a quadratic equation

$$y = 115.82+0.9812x + (-0.0081)*x^2 \text{ with } r^2=0.925 \quad (4.1)$$

which is a polynomial of second order. This polynomial shows the relationship of BP with age at the population level. We then drew the same trend line for each individual and found that it fit properly, i.e. the observed and prediction line were close to each other. This prediction demonstrated that quadratic regression fit BP quite well. Therefore, we used a simple quadratic regression model for each individual to obtain a better prediction and imputation

Figure 4.2 Observed mean of blood pressure by age, FHS



Before describing the basic steps in the imputation of BP, we imposed some limits on the prediction rules and on the number of consecutive imputations. To calculate the prediction equation, we excluded all values of BP up to 4 years before death and imputed missing BP where age was present (i.e. exam date is present but BP is missing). Imputation was only used for exams at which the subject turned up (i.e. age not missing). Theoretically, this method could be used to calculate the BP

value for any age of an individual regardless of whether they attended an exam or not.

We excluded the values just before death (in 4 years). The reason is that, at the end of life, the relationship between high blood pressure and mortality becomes less consistent, at which point BP is no longer predicted well by age. Thus trying to include these values would have made our prediction less accurate. These are therefore left out of our prediction derivation to avoid the problem.

Basic steps

Step 1 *Population level*

Plot the observed mean of BP against age and find out the shape of the observed curve for overall population.

Step 2 *Individual level*

(i)Regression coefficient: estimate regression coefficients for each individual. Let 'a' denotes regression constant, b and c regression coefficients. Using the value of a, b and c for each individual we made our prediction equation.

$$y_i = a_i + b_i * age_i + c_i * age_i * age_i \,, \tag{4.2}$$

where *i* is individual. Now for given value of age and parameters a, b and c, predict BP.

(ii) *Comparison*: plot the predicted BP with the observed one (Figures 4.2).

(ii) Imputation: for given value of age and parameters a, b and c, impute the missing values of BP.

The whole procedure is described with the help of an example. An individual entered the FHS at age 33, at which time his BP was 138 mm Hg. Except for the 2nd and 3rd exam, his BP was recorded up to 13th exam, after which he was absent from subsequent exams (i.e. his age was not recorded) and died between the 18th and 19th exam, which is recorded in 19th exam. We imputed the missing values from the 2nd and 3rd exam according to our algorithm. The variable BP was observed blood pressure and BP_4 was observed, but BP observations 4 years before death were excluded. The variables, BP_a, BP_b and BP_c represent regression (non-linear) parameter estimates of that individual. From the variable BPCODE, we can identify how many responses were missing and imputed. Code 1 indicates no missing and no imputation and code 2 is missing and imputed. The variable BPIMP shows imputed values using our approach.

Table 4.8 Blood pressure level of an individual before and after imputation, FHS

Exam	ID	Sex	Age	BP	Death	BP_4	BP_a	BP_b	BP_c	BPCODE	BPIMP
1	16	1	33	138	19	138	141.33	-0.50	0.01	1	138
2	16	1	36	missing	19	missing	141.33	-0.50	0.01	2	133.7
3	16	1	38	missing	19	missing	141.33	-0.50	0.01	2	133.9
4	16	1	39	140	19	140	141.33	-0.50	0.01	1	140
5	16	1	42	132	19	132	141.33	-0.50	0.01	1	132
6	16	1	43	122	19	122	141.33	-0.50	0.01	1	122
7	16	1	45	124	19	124	141.33	-0.50	0.01	1	124
8	16	1	47	134	19	134	141.33	-0.50	0.01	1	134
9	16	1	49	134	19	134	141.33	-0.50	0.01	1	134
10	16	1	51	150	19	150	141.33	-0.50	0.01	1	150
11	16	1	53	138	19	138	141.33	-0.50	0.01	1	138
12	16	1	56	170	19	170	141.33	-0.50	0.01	1	170
13	16	1	59	114	19	114	141.33	-0.50	0.01	1	114

4.5 Validation of the methods

To control for the adverse effects of missing data on a particular analysis, we made use of imputation. To assess the implications of the imputation method and associated assumptions, we needed a sensitivity analyses. Sensitivity analysis is used to ascertain how a given model output depends upon the input parameters. Two different algorithms (one for categorical variable and another one for continuous variable) are discussed and applied to impute the missing values in repeated measurement data. We assumed that the missing values in the FHS occurred at random. Thus, to validate our method and assumption, we performed a sensitivity analysis for the missing value imputation under different scenarios.

Overall, 26 percent of the missing responses in smoking status and 6 percent of the missing information on the BP risk factor were imputed. These imputed observations could influence the final outcome. Therefore, a sensitivity analysis of the imputation method was carried out. In the literature, most of the imputation methods deal with the mean value at population level. Using our method, missing values are imputed from the individual follow-up or observations that may not be comparable to other available methods. We carried out the sensitivity analysis based on the simple random selection procedures.

As illustrated, we imputed the missing values of two risk factors: smoking status and blood pressure, and created a scenario using simple random sampling technique for both the categorical and the continuous risk factors. We validated our imputation method based on this scenario. First, we imputed the missing values using the above method. Excluding the remaining missing observations after

imputation, we assumed a complete data set without any missing observations. Second, we deleted various observations randomly from this complete data set. In the scenarios, we deleted 10 percent, 20 percent and 30 percent of the observations based on simple random selection procedures. Third, we imputed the randomly deleted observations applying the same methods that we applied to impute the missing values for the original data set. Fourth, we compared the scenarios. We assumed that after imputation of the randomly deleted observations we would achieve the same output that was obtained after imputation for the original data set. If this assumption held, our methods were valid.

The scenarios for smoking status after imputing the randomly deleted responses are shown in Table 4.9. There were 61.9 percent non-smoking responses and 38.1 percent smoking responses in all recorded smoking statuses (excluding missing responses). After imputation, this distribution was 59.4 percent and 40.6 percent respectively. The change in the percentage distribution of imputed values after random deletion with different scenarios remains almost same. We found that even after deleting 30 percent of responses, the change in percent distribution of the number of responses was minimal. This nominal change was mainly due to the restriction that we imposed on the imputation of missing values, namely that the values for a maximum of two consecutive missing exams were permitted to be imputed.

Table 4.9 Scenarios for smoking status- responses are imputed after random deletion

Scenarios	Percent of non-smoking responses (number of responses)	Percent of smoking responses (number of responses)
Observed population	61.9 (33602)	38.1 (20687)
After imputation	59.4 (45507)	40.6 (31125)
Imputed after random deletion		
10%	59.2 (45303)	40.8 (31169)
20%	59.2 (44758)	40.8 (30891)
30%	59.3 (43779)	40.7 (30083)

The scenarios for blood pressure after imputing the randomly deleted responses are shown in Table 4.10 and Figure 4.3. The mean value of BP in the observed population was 136.87 mm Hg (SE 0.087). After imputation, this was 136.76 mm Hg (SE 0.0844). The change in the mean values of BP imputed after random deletion with different scenarios was almost negligible... Even after 30 percent random deletion, no substantial changes occurred in the BP mean and SE. After imputation with different scenarios, the total number of responses differed mainly because of two restrictions. (i) Imputation is only possible if we have at least

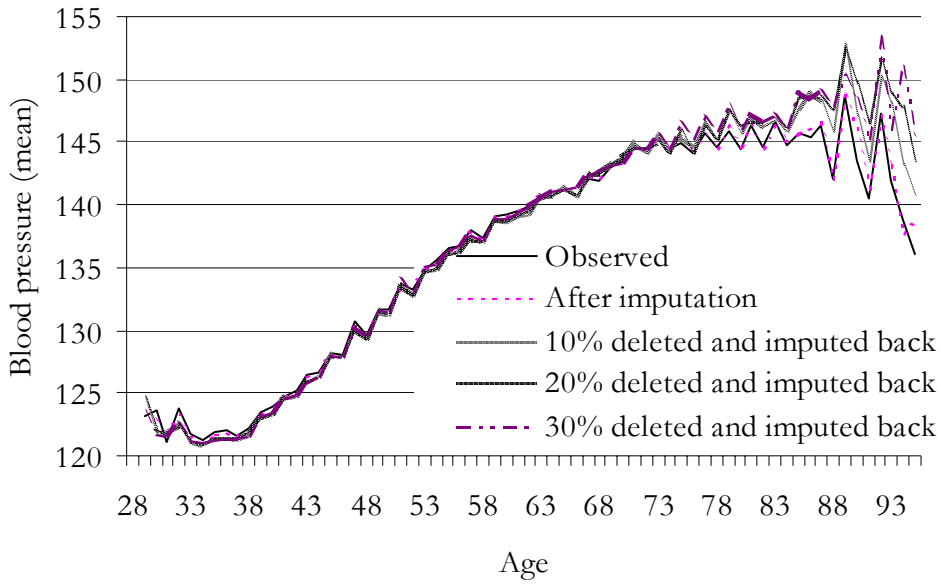
three responses. (ii) There is a problem of contiguity, i.e. values that are imputed too far away in time from the nearest time (i.e. age).

Table 4.10 Scenarios for blood pressure- responses are imputed after random deletion

Scenarios	BP mean	Standard error of mean	Total number of responses
Observed population	136.9	0.087	67496
After imputation	136.8	0.0844	71637
Imputed after random deletion			
10%	136.8	0.085	70079
20%	136.9	0.085	69872
30%	137.1	0.088	67189

The age-specific mean values of BP with different scenarios are presented in Figure 4.3. This figure shows that imputed age-specific BP means with different scenario remains almost unchanged. Estimated values with different scenarios differ slightly at older ages (after age 80), since at older ages, the relationship of BP with age is more complex, the number of responses is smaller and the number of missing observations is higher.

Figure 4.3 Scenarios of the age-specific mean values of blood pressure



4.6 Discussion

This chapter proposed to demonstrate a potential method to impute missing values of risk factors in a repeated measurement study. We illustrated the method using the well-known Framingham Heart Study. It was found that in over 48 years of follow-up, the percentage overall of non-missing observations was 41.47 percent in the Framingham Heart Study. We illustrated the missing values and the method of imputation used for two risk factors- smoking status as categorical variable and blood pressure as continuous variable, both of which were used in order to measure the risk career. The idea behind the proposed methodology is relevant to any study design subject to incomplete responses and having at least three repeated measurements.

We used a different method to impute missing records for a categorical risk factor than the method adopted for continuous risk factors. To impute smoking status, we followed the history of smoking status of the individual and imputed the value, which is closest in time to the true value, i.e. the value closest in time to the missing value. The method proposed for the imputation of BP is different. This method is an extension of regression procedure. Using the ordinary regression method, the expected population means are used to impute missing values. Using our regression method, we imputed the missing value for each individual from the information on each individual. The proposed method has several advantages over current approaches. Our method is easy to understand. The calculation procedure is very simple and not time consuming. Of paramount importance is that the curve has a good fit. These methods are easy to manipulate and applicable both for categorical and continuous risk factors in any long-time follow-up data. It also allows investigators to easily assess the sensitivity of the results obtained applying this method.

Using the proposed methods, we can easily reconstruct the risk factor career and measure the impact of this risk factor career from the life course perspective. For example, with the help of this method we were able to reconstruct the smoking career and to measure the life history of smokers and non-smokers, in order to construct the multistate smoking status life table in Chapter 5 of this study.

To validate our methods, we have presented the prevalence of smoking and blood pressure in the form of a scenario analysis. If we should wish to estimate the effect of imputation on some outcomes, e.g. CVD incidence and mortality, only the variability of the estimates will decrease because of the increasing sample observations. These observations are not independent of the observed values. Therefore, they should not be used. To predict the risk factor status at a specific age, we needed to impute the missing values. Using this approach, De Laet et al., (2003) predicted blood pressure at age 40, 50, and 60 and estimated the effect of

blood pressure on the life table outcomes. Based on the predicted blood pressure, the life table outcomes are more reliable than the observed information.

The major limitation of our method is that a certain number of rounds or waves are necessary, i.e. application of this method is possible only if at least three rounds of measurement are available. In developed countries, the number of longitudinal surveys has been on the rise for some time and the numbers of waves are increasing over time. Our method will be applicable to these data sets in the near future. Another limitation of this study is that if the individual returns to the study after a long period of time or the time span between one observation and the next is too long, the prediction may not fit with the observed trend.

Regarding the smoking career, we imputed missing values for two consecutive exams if bounded on both sides by non-missing values. Missing values have been imputed for BP only for the exam if the respondent's age was reported. Using the same approach, we may impute missing values for all exams and construct the full-length non-missing data set. We may conclude that by applying this method, it is possible to increase the utility of missing values in repeated measurement, and increase the statistical precision of final results.

References

Anderson KM, Odell PM, Wilson PWF, Kannel WB, (1991). Cardiovascular disease risk profiles. *American Heart Journal,* 121(1):293-298.

De Laet C, Peeters A, Mamun A, Bonneux L, (2003). Blood pressure evaluated during adulthood is a strong predictor for life expectancy and for life expectancy free of cardiovascular disease in both men and women. Manuscript, Erasmus University Rotterdam, The Netherlands.

Goldstein H, Woodhouse G, (1996). Efficient estimation with missing data in multilevel models. Paper was presented in the Eleventh International workshop on statistical modeling. Orvieto, Italy, 1996.

Hedeker D, Gibbons RD, (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods*, 2(1):64-78.

Heitjan DF, (1997). Annotation: What can be done about missing data? Approaches to imputation. *American Journal of Public Health*, 87(4): 548-550.

Laird NM, (1988). Missing data in longitudinal studies. *Statistics in Medicine,* 7:305-315

Little RJA, (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88:125-124.

Little RJA, Schenker N, (1995). Missing Data. In: Arminger, Clogg, and Sobel (eds.), *Handbook of Statistical Modeling for the Social and Behavioral Sciences.* New York: Plenum.

Little RJA, Rubin DA, (1987). Statistical Analysis with Missing Data. John Wiley and Sons, New York.

Little RJA, (1995). Modelling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90:1112-1122.

Little TD, Schabel KU, Baumert J, (2000). *Modelling Longitudinal and Multilevel Data: Practical Issues, Applied Approaches and Specific Examples*, Lawrence Erlbaum Associates, Mahwah.

Molenberghs G, (2001). Sensitivity analysis for incomplete data. Missing values: proceedings of a symposium on incomplete data. Proceedings of the tenth Symposium Statistical Software organized on November 8, 2001. Jaarbeurs Congress Centre, Utrecht, The Netherlands.

Rubin DR, (1976). Inference and missing data, *Biometrika*, 63:581-592.

Rubin DB, (1987*). Multiple Imputation for Nonresponse in Surveys*. John Wiley and Sons, New York.

Rubin DB, (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91: 473-489.

Schafer JL, (1997). *Analysis of Incomplete Multivariate Data*. Book number 72 in the Chapman and Hall series Monographs on Statistics and Applied Probability. Chapman and Hall, London.

Schafer JL, (2001). Multiple imputation in multivariate problems when the imputation and analysis models differ. Missing values: proceedings of a symposium on incomplete data. Proceedings of the tenth Symposium Statistical Software organized on November 8, 2001. Jaarbeurs Congress Centre, Utrecht, The Netherlands.

Shih WJ, (2002). Problems in dealing with missing and informative censoring in clinical trials, *Current Controlled Trials in Cardiovascular Medicine*, 3:1-7.

SOLAS 2.0, (1999). *Statistical Solutions – Versatile Methods for Data Aanalysis*. Use reference of SOLAS for missing data analysis 2.0.

Stamler J, Stamler R, Neaton JD, et. al., (1999). Low risk-factor profile and long-term cardiovascular and noncardiovascular mortality and life expectancy. *Journal of American Medical Association,* 282(21):2012-2018.

U.S. Census Bureau, (2002). Survey of income and program participation, SIPP data editing and imputation, http://www.sipp.census.gov/sipp, accessed September 13, 2002.

West CP, Dawson JD, (2002). Complete imputation of missing repeated categorical data: one-sample applications. *Statistics in Medicine*, 21(2):203-217.

Wei L, Shih JW, (2001). Partial imputation approach to analysis of repeated measurements with dependent drop-outs. *Statistics in Medicine*. 20:1197-1214.

William RM, (2000). Handling missing data in clinical trials: an overview. *Drug Information Journal*, 34:525-533.

Zhou XH, Eckert GJ, Tierney WM, (2001). Multiple imputation in public health research. *Statistics in Medicine*, 20:1541-1549.