An Introduction to Multiple Imputation Methods: Handling Missing Data with SAS® V8.2

Delfino Vargas-Chanes, PhD
Paul A. Decker, MS
Darrell R. Schroeder, MS
Kenneth P. Offord, MS

Technical Report #67
July 2003

TABLE OF CONTENTS

I. INTRODUCTION

Frequently data from survey research, epidemiology, or clinical settings because of unpreventable problems contain incomplete information caused by known or unknown reasons. In some situations it is possible to identify variables that are associated with the pattern of missing data. In other situations the mechanism of missing data is known but variables associated with this mechanism are inaccessible to the researcher. Finally, in some other situations the researcher controls the pattern of incomplete data (e.g. for an expensive clinical test the decision may be made to only perform the test on a random sample of study subjects).

The potential problem associated with missing data is that estimates from non-imputed data sets that exclude cases with at least one observation missing might not have enough power to detect significant differences and may introduce bias due to the reduced number of complete observations. More recently several alternatives have been proposed to solve this problem that included maximum likelihood [1] and Bayesian methods [2, 3].

This report is organized to give a general overview of the basic concepts of data imputation, with emphasis on application. The purpose is to explain the basic principles of multiple imputation for handling missing data and how to implement this method using SAS version 8.2.

II. BASIC CONCEPTS

One basic concept in data imputation is the mechanism of *ignorability*, the theoretical basis that explains the causes of missing data. Ignorability includes three complementary concepts missing completely at random (MCAR), missing at random (MAR), and non-ignorable (NI) missingness [4]. In the following paragraphs the concepts of MCAR and MAR are defined using formal

notation and examples to understand these fundamental concepts. NI mechanism is not the focus of this paper and has not been included.

Formal definitions

The matrix representation of a data set that includes observed and missing values is denoted by $Y=(Y_{obs}, Y_{mis})$, where Y_{obs} is the matrix of observed values, and Y_{mis} is the matrix of missing values. In this report we deal with recovering incomplete data to produce one (or many) data set(s) that can be used to estimate parameters for a specific analytic model. The complete probability density function of the data set is denoted as

$$p(Y \mid \theta) = \prod_{i=1}^{n} p(Y_i \mid \theta), \tag{1}$$

where θ denotes the parameters governing the distribution of Y. To better understand the missing data patterns, suppose that R is a matrix of indicators with the same dimensions as the original data matrix, where each element of R is 1 if the corresponding element in the original matrix is observed and 0 if it is missing. Since R has the same dimensions as the original matrix, the joint conditional probability is denoted as

$$p(Y, R \mid \theta, \phi) = p(Y \mid \theta) p(R \mid Y, \phi), \tag{2}$$

where ϕ denotes the conditional distribution of R given the complete data set Y. In expression (2) the complete data for the observed data is replaced and integrated over the missing portion and expressed as

$$p(Y_{obs}, R \mid \theta, \phi) = \int p(Y_{obs}, Y_{mis} \mid \theta) p(R \mid Y_{obs}, Y_{mis}, \phi) dY_{mis}. \tag{3}$$

The missing data mechanism is said to be "missing completely at random" (MCAR) if the distribution of the indicators R does not depend on either observed or missing data. Thus MCAR is defined as

$$p(R \mid Y_{obs}, Y_{mis}, \phi) = p(R \mid \phi), \tag{4}$$

In practical terms MCAR situation means that the mechanism that governs the missing data is not related either to the observed or missing data. The MCAR mechanism is equivalent to deleting a random subsample from a hypothetical population in which each observation has equal probability of being selected for deletion.

The second mechanism is "missing at random" (MAR). A formal definition of MAR states that the distribution of the missing data does not depend on the missing values but only on what we observe. Then

$$p(R \mid Y_{obs}, Y_{mis}, \phi) = p(R \mid Y_{obs}, \phi). \tag{5}$$

In other words, the missing data mechanism can be found in the data observed. Note that in this case the distribution of the observed values is not affected by incomplete information; only what is observed is relevant. Now substituting (5) into (3), we have

$$\begin{aligned} p(Y_{obs}, R \mid \theta, \phi) &= \int p(Y_{obs}, Y_{mis} \mid \theta), \phi) p(R \mid Y_{obs}, \phi) dY_{mis} \\ &= p(R \mid Y_{obs}, \phi) \times \int p(Y_{obs}, Y_{mis} \mid \theta), \phi) dY_{mis} \\ &= p(R \mid Y_{obs}, \phi) p(Y_{obs} \mid \theta). \end{aligned} \tag{6}$$

This means that if under MAR conditions the distribution of the parameter governing the missing data mechanism, ϕ, and the observed data, θ, are independent, then the joint distribution of the parameter space (θ, ϕ) can be split into the product of the parameter space θ and ϕ, a

property known as distinctness. The property if independence between θ and ϕ is useful when using maximum likelihood estimation.

From a practical standpoint MCAR scenarios are less common and the MAR mechanism is more frequent. For example, suppose for a study of cigarette smokers we collect information regarding number of cigarettes smoked daily and gender. If the probability of recording the number of cigarettes smoked daily does not depend on the number of cigarettes they smoke nor on gender then missing data are MCAR. On the other hand, if the probability of recording cigarettes smoked daily depends on gender, but not on number of cigarettes smoked, and this probability is the same across all subjects within gender, then data are MAR.

Empirical definitions

Let the $\mathbf{X}=(x_1,x_2, \ldots x_n)$ be observed variables, $\mathbf{Z}=(z_1,z_2, \ldots z_n)$ non observed covariates, and $\boldsymbol{Y}=(Y_{obs}, Y_{mis})$ the set of variables with missing values. In this case Y_{mis} are the variables to be imputed. Table 1 illustrates the MCAR, MAR patterns characterized by the observed and non-observed covariates.

MCAR patterns are very rare in the real world and survey researchers can take advantage of this pattern. This mechanism takes place when a random subsample of respondents is selected from a population to have complete information. For example, when a new survey instrument is developed (or a costly test is implemented) not all respondents are selected to answer that specific instrument (or complete that test) but only a random subsample. The missing data mechanism is MCAR and the imputation step is straightforward. In longitudinal studies the MCAR property would be useful if a random subsample are selected for follow up at each wave and every respondent has the same probability of being selected.

The MAR scenarios are more common in real practice. Under MAR assumption a set of covariates **X** is observed and the missing values, Y_{mis}, depend on the observed variables **X** (see Table 1). For example, if the number of cigarettes smoked by adolescents depends only on subject gender and number of peers who smokes tobacco and both variables are observed the missing data mechanism is MAR. There is no statistical test to prove this assumption; however, a common approach to see if MAR assumption is plausible is to determine if the covariates **X** are correlated with Y_{obs}, or **X** is associated with Y_{mis} (e.g. via logistic regression or chi-square test).

The imputation and the analytic models

A basic principle in data imputation under the MAR assumption is to understand the difference between the imputation model and the analytic model. The imputation model consists of all variables that collectively explain the missing data pattern and are useful in the imputation step. The analytic model is the one that is used to analyze the imputed data. Under the MCAR assumption, the variables used in the analytical model are, in general, the same as the variables of the imputation model (i.e. MCAR does not require covariates to explain patterns of incomplete data).

Under the MAR assumption identifying the variables for the imputation model becomes a relevant step to produce good quality imputations. The variables for the imputation model for the MAR condition are used to explain the missing data pattern and therefore are not necessarily restricted to the variables in the analytic model.

Another relevant question is how many and what variables in the imputation model are needed to provide good quality imputations. Under the MAR assumption the number and variables selected for the imputation model affect the quality of imputations as suggested in two studies [5, 6]. For example Graham and Schafer [5] showed that parameter estimates exhibit less

bias from population parameter estimates as the number of covariates included in the imputation model increase. However, a simulation study using data under a multivariate normal distribution assumption showed that better quality imputations are obtained when a subset of covariates are selected in the imputation model, particularly those associated with the missing data mechanism [6]. Therefore the strategy we will use in this report is to incorporate into the imputation model the variables that show some ability to predict the mechanism of missing data in two ways. First, using contingency tables or logistic regression models with categorical dummy variables for the missing data (to indicate whether the data are missing or not). Second, using a correlation matrix of the variables from the imputation model with the variables to be imputed. The analytic model may not include all the variables used in the imputation model; but all variables, which are to be used in the analytic model, need to be included in the imputation model. If the analytical model includes interaction terms, these interaction terms need to be included in the imputation step as well.

III. DATA AUGMENTATION

Multiple imputations (MI) incorporate a simulation process to fill-in several missing values since a single one might not reflect the variability. The variability results from the simulation process where missing data are filled after several iterations. MI are generated using MCMC methods from which several complete versions of the variables in the imputed data set are generated; each data set can be submitted to the analytical model using standard methods. A single set of parameter estimates is obtained from each model and Rubin's rules are used to pool all estimates into one (Appendix I).

One of the MCMC methods is called data augmentation (DA). DA is a particular method for dealing with missing data and has been described in detail in Tanner and colleagues [7, 8].

The general idea of DA via MCMC has two phases, called the I- and P- steps. The I- step: draw a random sample of observations from an initial marginal distribution in the first iteration

$$Y_{mis}^{(t+1)} \sim p(Y_{mis} \mid Y_{obs}, \theta^{(t)}).$$

The P-step: draw a random sample of parameters from a marginal distribution that incorporates observed and initial values for the missing observations from the I-step in the first iteration

$$\theta^{(t+1)} \sim p(\theta \mid Y_{obs}, Y_{mis}^{(t+1)}).$$

The I- and P- steps in the first iteration provide a starting value $\{Y_{mis}^{(0)}, \theta^{(0)}\}$ and posterior iterations create a stochastic Markov Chain of values $\{Y_{mis}^{(1)}, \theta^{(1)}; Y_{mis}^{(2)}, \theta^{(2)}; \ldots\}$ which converges in distribution to $P(\theta, Y_{mis}|Y_{obs})$. To produce multiple imputations we iterate over data augmentation and iterate over Y_{mis} to create a chain of filled in observations

$$\{Y_{mis}^{(t)}; Y_{mis}^{(2t)}, \ldots, Y_{mis}^{(mt)}\}.$$

This is equivalent to a run of m independent chains or burn in iterations of length t.

One important issue is to test the convergence of the MCMC process of a single chain and the number of iterations required for convergence depends on the amount of missing data and therefore varies from one data set to another. The algorithm requires assigning the minimum number of burn-in iterations, m, that are needed to guarantee that in a single chain θ^{t+m} will be independent from θ^{t}. This process can assure that after the burn-in period of size m every value estimated for θ can be taken as an independent draw from $P(\theta|Y_{obs})$ and Y_{mis} could be used as an imputed value.

Several methods have been proposed to investigate the convergence of the joint distribution of θ for a specified value m [9, 10]. From a practical approach the autocorrelation function (ACF) can be used to determine the convergence of the algorithm. For a lag-p stationary series $\{\kappa^{(t)} : t = 1,2,....k\}$ the ACF is defined as

$$\rho_p = \frac{Cov(\kappa^{(t)}, \kappa^{(t+p)})}{V(\kappa^{(t)})}. \quad (7)$$

The ACF plots r_p versus p for a limited number of p. The correlogram that is produced with the ACF plots helps to identify potential linear dependence. If the correlogram shows a sudden decay for the values p=2 to 4 this suggests serial independence, i.e. the algorithm converged to a satisfactory solution. The autocorrelation plots are easy to understand, however, it should be clear that they do not prove independence. Other methods are available to verify the convergence of the MCMC process but require further knowledge of Bayesian methods [11, 12].

IV. EXAMPLE: NICOTINE DEPENDENCE DATA

Data Source

A study of patients treated in the Mayo Clinic Nicotine Dependence Center (NDC) included 1877 adult patients that completed the Minnesota Multiphasic Personality Inventory (MMPI) prior to being treated for cigarette smoking. They were seen at the NDC between 4/1/1988 and 12/31/1999. Only patients with known smoking abstinence information 6-months following treatment were included in our analyses. At the time of the nicotine dependence consultation, patients were asked to complete an extensive baseline questionnaire. The questionnaire included demographic variables (e.g. age, gender, race, marital status, highest educational level) and tobacco use history (e.g. number of years smoked, number of cigarettes

smoked per day (CPD), longest duration of previous abstinence, type of tobacco products used). The questionnaire also contained items from the Fagerström Test for Nicotine Dependence (FTND), which assesses the severity of nicotine dependence [13]. At the time of the questionnaire administration, a nicotine dependence counselor evaluates and records the patient's stage of change according to the transtheoretical model [14]. As a routine part of the NDC follow-up a trained telephone interviewer, who is not associated with provision of the nicotine dependence intervention, attempts to contact patients by telephone 6-months following their NDC consultation. The patients' current tobacco use status is obtained as part of this call. For this study, tobacco abstinence is defined as a self-report of no use of any form of tobacco (not even a puff of a cigarette) during the previous 7-days.

The Analytic Model

The analytic model used for this investigation included specific personality characteristics (trait anxiety, depression, and neuroticism) as measured by the MMPI and whether they predict abstinence from tobacco 6-months following the nicotine dependence consultation. Five variables that have been found in previous investigations to be predictive of abstinence were included in the analytic model as covariates. Thus, the following expression is the logistic regression analytic model used in this example.

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + e_i, \tag{8}$$

where the dependent variable was the 7-day point prevalence tobacco use status 6-months post consultation (abstinence, y=1; use, y=0); the independent variables included pre-consult stage of change (x_1), average CPD smoked in the past six months prior to baseline consultation (x_2), longest duration without smoking (x_3), Fagerström score (x_4), gender (x_5, male=1, female=0) and

the selected MMPI scale of specific study interest (x_6). All patients had complete information on Y (6-months tobacco use status), x_5 (gender) and x_6 (MMPI scale). Note that separate analytic models were used for each MMPI scale of interest.

Of the 1877 patients, 42% (793) had complete data for model (8) the remaining had missing values for at least one adjusted variable. Since 58% of the patients were missing some information, there was a concern that those with complete data were not representative of the entire sample. To address this problem, MI methodology was proposed since this is the preferred method for analyzing data sets with missing data [2, 3, 15].

The Imputation Model

The data imputation methodology used in the example includes several steps. The first step was to determine the variables for the imputation model. The MAR assumption was used because we found this assumption plausible: there were a set of observed variables that were either correlated or showed an association with the variables with missing data. Second, we tested the convergence of the algorithm by using the ACF plots for the outcomes of interest in the imputation model. Finally the Rubin's rules, as described in Appendix I, were applied to summarize the parameter estimates for the analytic model.

For this example the MAR assumption was used in the imputation methodology because we hypothesized that some observed variables (hopefully with minimal amount of missing values) had some ability to explain the missing data mechanism.

Using the MAR assumption, it is necessary to identify the variables to be included in the imputation model. With binary indicator variables used to indicate missing data, the χ^2 test statistic was used to test the degree of association of the potential variables of the imputation

model with variables with missing data from the analytic model. In addition, Spearman rank correlation was used to estimate the association of the potential variables from the imputation model with the variables in the analytic model. Based on these analyses several variables were selected for inclusion in the imputation model. Tables 2 and 3 summarize the association with the variables in the imputation model.

In Table 2 the relationship between the variables in the imputation model and the variables in the analytic model is explored, some of the variables showed significant p-values from the χ^2 test as indicated. The following variables: age group, year of NDC consult, race, highest level of education, post-consultation stage of change, average CPD when smoking heaviest, longest duration without using tobacco, number of serious stop attempts, and location of appointment showed a significant association with at least one of variables from the analytic model.

In Table 3 the variables post-consultation state of change, average CPD currently, average CPD when smoking heaviest, and longest prior duration without using tobacco at time of consultation, showed a significant correlation with the variables from the analytic model.

Programming Multiple Imputation

In this section the actual implementation of the imputation procedure using SAS version 8.2 is explained. After selecting the variables for the imputation model as described above we prepared the SAS statements using PROC MI. This procedure assumes that the variables included in the statements belong to the analytic and imputation models. The SAS statements used in this report are as follows:

```
(line)
(1)   proc mi data=final out=miout nimpute=8 seed=101039
(2)   round=1
(3)   minimum=2 0 1988 0 0 0 0 1 1 1 1 1 1 1 1 0 0 0 0 0 0 .
(4)   maximum=8 1 1999 1 1 1 1 8 3 3 4 4 4 3 3 1 4 1 1 1 1 .;
(5)   mcmc initial=em prior=ridge=0.5 outiter =outit
(6)        timeplot (mean(pre_con avg6mos2 longstop ftndgrp))
(7)        acfplot  (mean(pre_con avg6mos2 longstop ftndgrp));
(8)   var  agegrp gender ndcyear race_new single together divorce
(9)        educ pre_con post_con avgcurr2 avg6mos2 avgheav2
(10)       longstop longest ftndgrp numstop loc_clin hospital ndc
(11)       success6 tp_0 tp_1 tp_2 tp_3 tp_4 tp_5 tp_6 tp_7 tp_8
(12)       tp_9 tp_f tp_k ts2_1 ts2_4 ts2_7 ts2_8 ts2_9 t_t a
(13)       lw1_231 dl j09 depw j01 p21 pk1 p84 _5ne hr1 hr2 hr3
(14)       man psm smo;
(15)  run;
```

The MI Procedure.- The SAS code in the proc statement (line 1) instructs the program to read and save SAS files. In addition, in this line the user specifies more imputation details like the number of imputations, a random seed, rounding numbers and minimum and maximum values for each variable.

The option "`data=final`" refers to a matrix where the original data are stored, each row corresponds to one patient and the columns correspond to the variables. The incomplete data are defined as missing values and PROC MI will impute a value for each missing observation. An output file is produced with the statement "`out=miout`", this file contains all 8 imputed datasets with the missing values replaced with imputed values.

To determine the number of imputations needed we use formula 9 (Appendix I). In this study we expect about 58% of the observations will be missing for one or more variables in the analytic model (8). Then for k=8 imputed data sets the expected relative efficiency for recovering missing values will be close to 93% (i.e. $(1+0.58/8)^{-1}\times 100$). Thus the option "`nimpute=8`" will give the desired relative efficiency in estimating the parameters for the analytic model in expression (8) using the multiple imputation methodology. The seed is a

random number and it is specified as `seed=101039` and is used mainly to reproduce the simulation based on the same seed. If the `seed` statement is omitted then in each run the seed number will be randomly selected.

The option "`round=1`" (line 2) results in all variables in the imputed data sets to be integers. The "`minimum=`" and "`maximum=`" lines (3 and 4) specify the minimum and maximum values for all the variables in order of the "`var`" statement (lines 8 to 14). The value of "•" in this statement implies that the variable has no bounds. The variables in the `var` statement are in the same order as in the `round` statement. This alternative provides more flexibility allowing the simulation process to converge in fewer iterations.

The round option can also be set to specific formats for each variable. For example, "`round=.01`" indicates rounding all variables to the nearest multiple of .01. The option "`round=• 1 10`" indicates no rounding for the first variable, rounding to the integer for the second and to the nearest 10 for the third.

The MCMC options.- The options set in line (5) indicate the Markov Chain Monte Carlo method indicate the starting values, the type of prior distribution to be used, and the output for the autocorrelation plots. In this example the MCMC method has been chosen with an initial estimation provided by the EM algorithm. Running this algorithm as the initial step in the data imputation process is highly recommended since the initial estimates obtained from the EM-algorithm are used as starting values for the simulation step.

In this example the Ridge method was selected as the prior option. The Ridge prior option is useful when the estimated covariance matrix is nearly singular. It is the option of choice when the variables in the imputation data set could have a high correlation, or the number

of rows is substantially less than the number of columns, or when some of the variables show little variability. Since in our example the expected level of correlation could be high we selected the "`prior=ridge=0.50`" as an option to make the algorithm converge efficiently. The number 0.50 indicates the proportion of the data set that is to be used to generate the initial estimates.

The option "`prior=jeffreys`" is given by default and uses a non-informative prior. The non-informative prior starts with no information available for the parameter of interest; using MCMC methods it generates a posterior distribution and generates updated parameters; the process iterates until a convergence criterion is met. In general Jeffrey's method is quite stable and works for most data imputation situations.

The option "`prior=input=<sas dataset>`" specifies the informative prior where the mean and covariance is requested with this statement. The starting values, in the informative prior method, are chosen from an *imaginary* dataset. Instead of choosing an informative prior starting with complete ignorance of the parameter of interest, as the Jeffreys methods does, an estimated mean and covariance matrix is obtained after observing a sample from the data. Thus, the process for generating the updated parameters creates an *imaginary* prior distributions and represents the best guess of the population covariance. A similar iterative process as in the Bayesian approach generates the final estimates until a convergence criterion is met. The option "`outiter`" creates a SAS output dataset that contains all relevant information for each imputation.

The default statements for this procedure are the number of burn-in iterations. This option refers to the number of iterations required before stopping the first chain of simulations, as described in section III Data Augmentation in this report. The defaults given by SAS are

```
chain=single nbiter=200 niter=100 initial=em.
```

There has been no particular preference whether single or multiple chains would provide the best results. Is running a single run of size *mt* (i.e. 200×100) better than having *t* (i.e. 200) parallel runs of size *t* (i.e. 100)? It is open to debate, the default option in SAS is single chain and S+ recommend to run multiple chains [16]. See section III Data Augmentation regarding the burn in iterations concept.

The parameter "`nbiter=200`" specifies the number of burn-in iterations within each chain and `niter=100` specifies the number of iteration between imputations in a particular chain. When a single chain is chosen there will be 200×100 iterations. The options "`nbiter=200`" and `niter=100` are selected as the default since the values are quite reasonable for the current data set.

The Variables in the Imputation Model.- The SAS statement that includes the variables from the imputation and the analytical models are specified in lines (8) to (14), that are the main interest of the study. However, we have included more variables since other models are of interest as well for further analyses.

Testing the Convergence

An important step in data imputation is to test the convergence of the MCMC process. This step is recommended after specifying the number of iterations (`burn-in` and `niter`), and the number of datasets to be imputed (`nimpute=`). Lines (6) and (7) request the autocorrelation function (ACF) plots for the outcomes in the analytical model. As discussed in a previous section, the ACF plots are useful to examine the dependency of the time series to diagnose convergence of the MCMC process.

Figures 1 to 4 show the ACF for variables `pre_con` (pre-consult stage of change), `avg6mos2` (average CPD past 6 months), `longstop` (longest duration without smoking), ftndgrp (Fagerström score) in the analytic model as requested in lines (6) and (7). Figure 1 shows a slow decay for the first four autocorrelations, meanwhile Figures 2 to 4 show a sudden decay of the stationary process. Overall the ACF functions in Figures 1 to 4 provide some evidence that the stationary process converged to a satisfactory solution.

Rules for Reporting Final Results

A final step for reporting the results is to apply Rubin's rules to the estimates from the imputed data sets. With eight data sets we conduct logistic regression analysis on each data set. The "`proc logistic`" statements that request a logistic regression model by imputation are:

```
proc logistic data=final covout outest=out1 noprint;
  by _imputation_;
  model success6= t_t gender contemp prepare cpd21_39 cdp40 ftnd_6
  long1_30 long1;
```

Independent variables in this model included TSC tension scale (t_t), gender (with values 1=male, 0=female), dummy variables for stage of change (contemp=1 if in contemplation stage and 0 otherwise; and prepare=1 if in preparation/action stage, 0 otherwise), dummy variables for average cigarettes in the past 6 months (cpd21_39 = 1 if cpd = 21-39 or 0 otherwise; cdp40=1 if cpd $\geq$ 40, 0 otherwise), Fagerström score (ftnd_6 = 1 if score $\geq$ 6, 0 otherwise), and dummy variables for longest duration without smoking (longl_30 if duration = 1-30 days, and long1 if duration is more than 1 month). This "`proc`" statement will produce 8 parameter estimates for each of the coefficients in the analytic model, but will not be printed as specified by the "`noprint`" option. Instead all parameter estimates will be stored in the "`out1`" data set. This data set will have the 9 parameter estimates corresponding to the logistic regression model for

each imputation. In this example there were 8 models with 9 parameters in each model (data not shown).

Equations (1) to (7) from Appendix I are applied to the eight models to generate the output for the MIANALYZE procedure. The PROC MIANALYZE statements in SAS are:

```
proc mianalyze data=out1 edf=1867;
  var intercept t_t gender contemp prepare cpd21_39 cdp40 ftnd_6
  long1_30 long1;
```

Output 1 shows the results of applying Rubin's rules, see Appendix I. The variance between estimates (column 2) corresponds to equation (3, Appendix I); the variance within estimates (column 3) are related to equation (2, Appendix I); the square term of expression (4, Appendix I) gives the total variance (column 4); the degrees of freedom (column 5) are calculated using equation (5, Appendix I); the relative increase in variance (column 6) corresponds to equation (8, Appendix I); the fraction of missing information (column 7) corresponds to expression (7).

Output 1. Rubin's Rules using PROC MIANALYZE

		(Column)				
(1)	(2)	(3)	(4)	(5)	(6)	(7)
	----------Variance----------					
Variable	Between	Within	Total	DF	Relative Increase in Variance	Fraction Missing Information
Intercept	0.017182	0.070552	0.089881	137.17	0.273981	0.225230
t_t	0.000001155	0.000054809	0.000056109	1598.5	0.023717	0.023317
gender	0.000086216	0.011086	0.011183	1812.8	0.008749	0.008695
contemp	0.013492	0.050266	0.065444	119.3	0.301963	0.243468
prepare	0.021500	0.050227	0.074415	62.942	0.481575	0.344534
cpd21_39	0.003470	0.014484	0.018387	140.46	0.269521	0.222253
cpd40	0.011829	0.031332	0.044640	74.292	0.424744	0.315288
ftnd_6	0.008979	0.013436	0.023538	36.697	0.751789	0.456996
long1_30	0.007617	0.018912	0.027481	68.171	0.453098	0.330168
long1	0.005917	0.022781	0.029437	125.04	0.292197	0.237187

From Output 1 the relative efficiency (RE) can be calculated using equation 9 (Appendix I). RE reflects the efficiency of the imputation process after k=8 imputations, not the efficiency of the parameter estimates. For example, RE for the variable ftnd _6 is equal to 0.95[a] which means that the efficiency of the imputation process for this variable is 95%[b]. The percentage of relative efficiency (in parenthesis) of the remaining variables in the logistic model are t_t (100%), gender (100%), contemp (97%), prepare (96%), cpd21_39 (97%), cpd40 (96%), long1_30 (96%), long1 (97%).

The 95% Confidence Intervals (CI) and the significance level for the parameter estimates from the logistic regression are provided in Output 2. They are obtained from eight complete data sets and summarized using Rubin's Rules (Appendix I).

PROC MIANALYZE supports other models that apply Rubin's Rules as presented in Appendix I, see SAS Technical report 8.2 for further reference on the procedures that are supported by multiple imputation analysis [17].

Finally, Output 3 presents the logistic regression model with no imputed data. The sample size is reduced to 793 observations, due to missing values. By comparing Outputs 2 and 3 and some similarities and differences emerge. First, notice that standard errors in Output 2 are smaller than in Output 3. For example, the standard errors for gender are 0.11 and 0.16 for imputed and non imputed data, respectively. The reason being is that for the imputed data we have more observations available.

---

[a] using the fraction of missing information (column 7 from Output1), RE is calculated as $(1+0.457/8)^{-1}$

[b] Note that for variables with RE equal to 100% there are no missing data or it is less than 5% thus the corresponding parameter estimate is close to the actual estimate.

Output 2. Parameter estimates and 95% Confidence Intervals using PROC MIANALYZE

The MI Procedure
Multiple Imputation Parameter Estimates

Variable	Mean	Std Error	95% Confidence Limits	
Intercept	-1.456105	0.299802	-2.04894	-0.86327
t_t	-0.020665	0.007491	-0.03536	-0.00597
gender	0.405855	0.105749	0.19845	0.61326
contemp	0.500577	0.255820	-0.00596	1.00711
prepare	0.878217	0.272791	0.33308	1.42336
cpd21_39	-0.287429	0.135600	-0.55551	-0.01935
cpd40	-0.640260	0.211281	-1.06122	-0.21930
ftnd_6	-0.099658	0.153420	-0.41060	0.21129
long1_30	0.337200	0.165775	0.00642	0.66798
long1	0.514874	0.171573	0.17531	0.85444

Multiple Imputation Parameter Estimates

Parameter	DF	Minimum	Maximum
Intercept	137.17	-1.711285	-1.315350
t_t	1598.5	-0.022588	-0.018962
gender	1812.8	0.393348	0.419131
contemp	119.3	0.336356	0.671839
prepare	62.942	0.638929	1.084117
cpd21_39	140.46	-0.356567	-0.197010
cpd40	74.292	-0.815684	-0.454293
ftnd_6	36.697	-0.235057	0.040653
long1_30	68.171	0.181206	0.449794
long1	125.04	0.365602	0.594621

Multiple Imputation Parameter Estimates

Parameter	Theta0	t for H0: Parameter=Theta0	Pr > |t|
Intercept	0	-4.86	<.0001
t_t	0	-2.76	0.0059
gender	0	3.84	0.0001
contemp	0	1.96	0.0527
prepare	0	3.22	0.0020
cpd21_39	0	-2.12	0.0358
cpd40	0	-3.03	0.0034
ftnd_6	0	-0.65	0.5200
long1_30	0	2.03	0.0458
long1	0	3.00	0.0033

Second, the estimates are in the correct direction in both models. For example, the estimates for the variables T_T, CPD21_39, CPD40, and FTND_6, are negative in Outputs 2 and

3. Third, the estimates from the non imputed data (Output 3) are included in the 95%CI limits for the estimates from the imputed data.

Output 3. Logistic regression with non imputed data, N=793.

The LOGISTIC Procedure
Model Information

Data Set	WORK.FINAL
Response Variable	SUCCESS6 Tobacco abstinence (6 month)
Number of Response Levels	2
Number of Observations	793
Model	binary logit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	SUCCESS6	Total Frequency
1	Abstinent	265
2	Using tobacco	528

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.7870	0.4545	15.4621	<.0001
T_T	1	-0.0116	0.0111	1.0971	0.2949
GENDER	1	0.3543	0.1597	4.9227	0.0265
CONTEMP	1	0.9088	0.4019	5.1143	0.0237
PREPARE	1	1.3704	0.4013	11.6607	0.0006
CPD21_39	1	-0.1939	0.2001	0.9396	0.3324
CPD40	1	-0.6697	0.2732	6.0086	0.0142
FTND_6	1	-0.1485	0.1915	0.6010	0.4382
LONG1_30	1	0.2154	0.2085	1.0667	0.3017
LONG1	1	0.4040	0.2236	3.2634	0.0708

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
T_T	0.988	0.967	1.010
GENDER	1.425	1.042	1.949
CONTEMP	2.481	1.129	5.455
PREPARE	3.937	1.793	8.644
CPD21_39	0.824	0.557	1.219
CPD40	0.512	0.300	0.874
FTND_6	0.862	0.592	1.255
LONG1_30	1.240	0.824	1.867
LONG1	1.498	0.966	2.322

Fourth, all variables but five (e.g. t_t, cpd21_39, ftnd_6, long1_30, and long1) were statistically significant ($p<0.05$) in Output 3; however, all variables but two (e.g. contemp and ftnd_6) was statistically significant ($p<0.05$) in the imputed data set in Output 2.

V. DISCUSSION

In practice we face the dilemma of making inferences restricting the data set to those with non-imputed data, or using imputed data. In either scenario we are making assumptions. Estimates are potentially biased since the assumption made is that data are missing completely at random and it is unlikely that this assumption is tenable.

Having 10% or less of missing data does not merit the effort of conducting imputations. Inferences are reliable for both the non-imputed and imputed data. With more than 10% of incomplete information there is generally some concern that the inferences based on complete data are not warranted. The first assumption that we want to make is that data are MAR. Since, in most practical situations incomplete data does not occur completely at random instead we assume MAR and rely on observed variables that are related to and hence can predict imputed values for the missing data.

The MAR assumption appears to be plausible in our example since, in the analyses done, variables were identified to be associated with the missing value status of the variables in the analytic model (Tables 2 and 3). However, currently there is no statistical test available to test whether data are MAR. One limitation is that variables included in the imputation model are restricted to those available in the data set. We make the theoretical assumption that those variables are sufficient for the MAR condition. Researchers should have some theoretical

understanding that the observed variables are predictive of the missing value mechanism associated with the MAR assumption.

In our example, multiple imputation demonstrated an efficiency ranging from 93 to 100% for recovering missing values, which is considered satisfactory. Some discrepancies are expected between estimates obtained using imputed data compared to non-imputed data, as observed in Outputs 2 and 3. The main reason is that if data are MAR there exist particular characteristics associated with subjects with incomplete information that distinguish them from subjects with complete information. Moreover, having more complete observations available decreases the standard error; however, the MI method penalizes for the simulation involved in recovering missing data by adjusting the standard errors. Since Rubin's Rules correct for the amount of missing data, standard errors are inflated relative to what would be obtained if all data were observed. However, in all cases the adjusted standard errors obtained from the MI method are smaller than those obtained if the analysis is restricted to the subset of subjects with non-imputed data.

Simulation studies demonstrate that multiple imputations are robust and provide satisfactory results [18, 19]. The simulation studies suggest that even if the MAR assumption cannot be totally defended, imputation methods provide less biased estimates as compared to analyses using complete data only[3]. This is because making incorrect assumptions in the model only affects the portion of the data that was imputed and therefore the influence on the parameter estimates is lessened by the non-imputed data [3].

Some caution needs to be exercised when conducting MI. The multiple imputation methodology is not able to recover data from a bad study design. If there is a concern that missing data happened because of a bad study design, MI methods are not warranted for

successful recovery of incomplete information. Furthermore, MI methods are not designed for making a forecast of missing data at the individual level. Instead, MI methods are intended for producing simulated data that preserves the structure from the available data, based on its prior distribution, and its covariance matrix. The imputed data can be used for making inferences for the combined estimates.

Finally, in some cases the missing data mechanism depends on the variable itself and is called Non-Ignorable (NI) missing. For example, if the probability of recording cigarettes smoked daily depends only on the number of cigarettes smoked daily (i.e. it does not depend on other variables available in the data set or possibly it depends on variables that were not recorded), then the mechanism of missing data is NI. The NI scenario is the most difficult to analyze and is under current investigation. Fairclough and colleagues [20, 21] have addressed this problems using patterned mixed models. Other more complex modeling strategies include the use of Bayesian methods [11].

PROC MI provides satisfactory results when data support the normality assumption and when the missing data mechanism is either MCAR or MAR; however, it is not recommended for NI missing data patterns.

MI is a method for imputing data under MAR or MCAR assumptions using MCMC methods and once the data set has all missing values recovered, regular modeling can be applied to each data set and later the estimates need to be combined for obtaining the final estimates to be reported. PROC MI and PROC MIANALYZE as implemented in SAS version 8.2 are useful tools for data imputation.

VI. REFERENCES

1. Little, R.J.A., *Regression with missing X's: A review.* Journal of the American Statistical Association, 1992. **87**: p. 1227-1237.
2. Rubin, D.B., *Multiple imputation for nonresponse in surveys.* 1987, New York, USA: John Willey & Sons.
3. Schafer, J.L., *Analysis of incomplete multivariate data.* 1997, New York: Chapman Hall.
4. Rubin, D.B., *Inferences with missing data.* Biometrika, 1976. **63**(3): p. 581-592.
5. Graham, J.W. and J.L. Schafer, *On the performance of multiple imputation for multivariate data with small sample*, in *Statistical strategies for small sample research*, R.H. Hoyle, Editor. 1999, SAGE publications, Inc.: Thousands Oaks, CA. p. 1-29.
6. Russell, D.W., H.S. Stern, and S. Sinharay, *An evaluation of multiple imputation as an approach to missing data.* 2000.
7. Tanner, M.A. and W.H. Wong, *The calculation of posterior distributions by data augmentation (with discussion).* Journal of the American Statistical Association, 1987. **82**: p. 528-550.
8. Tanner, M., *Tools for statistical inference, methods for the exploration of posterior distributions and likelihood functions.* 1993, New York: Springer-Verlag.
9. Ritter, C. and M.A. Tanner, *The Gibbs stopper and the griddy Gibbs sampler.* Journal of the American Statistical Association, 1992. **87**: p. 861-868.
10. Roberts, G.O., *Convergence diagnosis of the Gibbs sampler*, in *Bayesian Statistics*, J.M. Bernardo, et al., Editors. 1992: Oxford University Press.
11. Gilks, W.R., S. Richardson, and D.J. Spiegelhalter, *Markov Chain Monte Carlo in Practice.* 1996, Washington, D.C.: Chapman & Hall/CRC. 486.
12. Casella, G. and R.L. Berger, *Statistical Inference.* Second ed. 2001, Belmont, CA: Duxbury Advanced Series.
13. Heatherton, T.F., et al., *The Fagerström test for nicotine dependence: A revision of the Fagerström Tolerance Questionnaire.* British Journal of Addiction, 1991. **86**: p. 1119-1127.
14. Prochaska, J.O. and C.C. DiClemente, *Stages and processes of self-change of smoking: Toward an integrative model of change.* Journal of Consulting and Clinical Psychology, 1983. **51**: p. 390-395.
15. Hall, S.M., et al., *Statistical analysis of randomized trials in tobacco treatment: longitudinal designs with dichotomous outcome.* Nicotine and Tobacco Research, 2001. **3**: p. 193-202.
16. Splus, *Guide to Statistical and Mathematical Analysis, Version 4.0.* Vol. 2000. 2000, Seattle: StatSci, a division of MathSoft, Inc.
17. SAS Institute Inc, *SAS/STAT Software: Changes and Enhancements. Release 8.2.* 2001, Cary, NC: SAS Institute Inc.
18. Ezatti-Rice, T.M., et al. *A simulation study to evaluate the performance of model-based multiple imputations in NCHS health examinations surveys.* in

Proceedings of the Annual Research Conference. 1995. Washignton D.C.: Bureau of the Census.
19. Vargas-Chanes, D., *Imputation methods for incomplete panel data with applications to latent growth curves*, in *Dept. of Sociology*. 2000, Iowa State University: Ames, Iowa. p. 99.
20. Troxel, A.B., et al., *Statistical analysis of quality of life with missing data in cancer clinical trials.* Stat Med, 1998. **17**(5-7): p. 653-66.
21. Fairclough, D.L., et al., *Comparison of several model-based methods for analyzing incomplete quality of life data in cancer clinical trials.* Stat Med, 1998. **17**(5-7): p. 781-96.
22. Schafer, J.L. and M.K. Olsen, *Multiple imputation for multivariate missing-data problems: A data analyst's perspective.* Multivariate Behavioral Research, 1998. **33**(4): p. 545-571.

APPENDIX I.

Rubin's Rules

For the k sets of imputed values, let $\hat{\theta}_l$, l=1,..., k, denote the parameter estimates obtained from the analytic model each for complete dataset. The combined estimate of θ is $\overline{\theta}$, the average of the parameter estimates over the k-data sets. Namely,

$$\overline{\theta} = \frac{1}{k}\sum_{i=1}^{k}\hat{\theta}_i . \tag{1}$$

Let $\hat{W}_i$ represent the standard error associated with parameter estimate $\hat{\theta}_l$, l=1,..., k, the within-imputation variance is

$$\overline{W} = \frac{1}{k}\sum_{i=1}^{k}\hat{W}_i^2 . \tag{2}$$

the between-imputations variance in $\hat{\theta}$ is estimated as

$$B = \frac{1}{k-1}\sum_{i=1}^{k}(\hat{\theta}_i - \overline{\theta})^2 , \tag{3}$$

thus the total standard deviation associated with $\overline{\theta}$ is

$$\sqrt{T} = \sqrt{\overline{W} + \frac{k+1}{k}B} , \tag{4}$$

that can be used to estimate a confidence interval for $\overline{\theta}$ with *df* degrees of freedom. Using a t-Student distribution with degrees of freedom based on Rubin's [2] formula

$$df = (k-1)\left(1 + \frac{k}{k+1}\frac{\overline{W}}{B}\right)^2 . \tag{5}$$

Thus, a 100(1-α)% interval estimate for θ is

$$\bar{\theta} \pm t_{df,1-\alpha/2}\sqrt{T}\,. \tag{6}$$

The multiple imputation efficiency can be calculated using Rubin's [2] formula to estimate the fraction of missing information about $\bar{\theta}$ as

$$\lambda = \frac{r + 2/(df+3)}{r+1}, \tag{7}$$

where

$$r = \frac{(1+k^{-1})B}{W} \tag{8}$$

The ratio r is known as the relative increase in variance due to nonresponse [2]. Thus the relative efficiency (RE) due to *k* imputations expressed in terms of the variance is

$$RE = \left(1 + \frac{\lambda}{k}\right)^{-1}. \tag{9}$$

Table 4 depicts some values for *k* and λ that can be used for determining the efficiency of data imputation. For example, the percentage of efficiency in recovering missing data with 30 percent missing, with five (*k*=5) datasets, is 94 percent. A general rule of thumb is that 3 to 5 imputations are sufficient to obtain good quality overall estimates [3, 22].

Tables and Figures

Table 1. Characterization of MCAR, and MAR patterns

Table 2: Assessing the association between the missing value status of variables in the analytic model (columns) and candidate predictor variables (rows) for missing data in the imputation model

Table 3: Spearman rank correlations coefficients between variables in the analytic model (columns) and candidate predictor variables (rows)

Table 4. Efficiency expressed as a percent for multiple imputations (MI) with k data sets and λ percent missing information. Rubin [2]

Figure 1. Autocorrelation Function (ACF) plot for x_1:pre-consult stage of change (PRE_CON) variable with 21 iterations shown. The ACF plot decays after 5 iterations, 95% CI intervals are shown with dotted lines

Figure 2. ACF plot for x_2:average CPD past 6 months (AVGMOS2) variable. The ACF decays after the second iteration, 95% CI intervals are shown with dotted lines

Figure 3. ACF plot for x_3: Longest duration without smoking (LONGSTOP) variable. The ACF decays after the fourth iteration, 95% CI intervals are shown with dotted lines

Figure 4. ACF plot for x_4: Fagerström score (FTNDGRP) variable. The ACF decays after the second iteration, 95% CI intervals are shown with dotted lines

Table 1. Characterization of MCAR, and MAR patterns

Patterns	**Covariates X=(x_1,x_2, ...x_n) are observed**	**Covariates Z=(z_1,z_2, ...z_n) are not observed**
MCAR	Y_{mis} does not depend on **X**	Y_{mis} does not depend on **Z**
MAR	Y_{mis} does depend on **X**	Y_{mis} does not depend on **Z**

Table 2: Assessing the association between the missing value status of variables in the analytic model (columns) and candidate predictor variables (rows) for missing data in the imputation model

	Variables in the analytic model with missing values			
Candidate predictor variables for the imputation model	**x_1: Pre-consult stage of change**	**x_2: Average CPD past 6 months**	**x_3: Longest duration without smoking**	**x_4: Fagerström Score**
	% † [P ‡]	% † [P ‡]	% † [P ‡]	% † [P ‡]
x_5:Gender	[NS]	[NS]	[NS]	[NS]
Female	44.3	28.8	33.5	31.0
Male	41.3	31.1	33.2	33.0
x_6:Year of NDC consult	[<0.001]	[<0.001]	[<0.001]	[<0.001]
1988	100.0	16.0	22.2	13.8
1989	98.5	9.7	10.1	9.0
1990	58.4	37.5	41.2	38.3
1991	34.6	37.3	37.3	35.9
1992	43.5	45.9	45.9	44.5
1993	29.3	32.5	40.1	35.0
1994	29.0	33.3	41.2	31.6
1995	31.0	47.1	48.3	42.5
1996	19.0	28.5	34.5	31.0
1997	7.3	42.4	46.1	52.7
1998	1.0	6.1	16.3	22.5
1999	0.9	9.4	15.0	21.5
x_7: Age group (years)	[0.005]	[NS]	[NS]	[NS]
18-29	45.2	28.9	28.9	27.9
30-39	50.8	31.7	36.3	34.2
40-49	40.1	28.5	31.0	31.8
50-59	42.3	29.4	33.7	30.0
60-69	43.3	32.1	33.8	32.8
70-79	30.8	24.2	36.3	33.0
80-89	16.7	33.3	33.3	50.0
x_8: Race	[0.036]	[NS]	[NS]	[NS]
Non-Caucasian	18.2	9.1	12.1	15.2
Caucasian	35.9	17.3	20.9	18.7
x_9: Marital Status	[NS]	[0.101]	[0.101]	[0.080]
Separated/divorced/widowed	44.7	33.3	37.2	35.8
Single	36.7	26.0	29.0	27.2
Married/living together	42.9	28.4	32.1	30.6
x_{10}: Highest level of education	[<0.001]	[0.022]	[NS]	[NS]
< 8th grade	12.3	6.2	10.8	7.7
8th grade	16.2	7.5	8.7	2.8
Some high school	32.9	2.9	9.0	4.9
High school	39.4	3.1	5.4	5.1
Some college	32.7	1.9	5.8	2.9
≥4 year degree	35.1	1.8	3.5	7.0

	Variables in the analytic model with missing values			
Candidate predictor variables for the imputation model	**x_1: Pre-consult stage of change**	**x_2: Average CPD past 6 months**	**x_3: Longest duration without smoking**	**x_4: Fagerström Score**
	% † [P ‡]	% † [P ‡]	% † [P ‡]	% † [P ‡]
x_{11}: Post-consultation stage of change	[0.045]	[0.004]	[NS]	[0.016]
Pre-contemplation	0.0	16.7	20.0	16.7
Contemplation	2.2	6.4	15.9	10.7
Preparation/action	0.5	14.4	18.8	18.8
x_{12}: Average CPD when smoking heaviest	[0.006]	[0.097]	[NS]	[NS]
1-20	23.3	4.7	9.1	5.9
21-39	28.5	1.8	6.8	7.5
≥40	34.1	4.2	10.2	5.8
x_{13}:Average CPD currently smoking	[NS]	[NS]	[NS]	[NS]
None	21.1	0.0	21.1	0.0
1-20	30.3	4.6	9.8	7.9
21-39	26.5	4.0	8.5	7.7
≥40	32.5	3.1	10.8	5.2
x_{14}: Longest duration without using tobacco	[NS]	[0.068]	[0.002]	[NS]
< 1 day/not at all	27.4	6.3	12.6	6.8
1-30 days	31.7	2.8	7.2	4.8
> 1 month	29.4	4.8	5.0	4.5
x_{15}: Number of serious stop attempts	[0.007]	[NS]	[0.043]	[NS]
None	29.2	5.3	0.0	10.6
1	22.8	2.9	5.8	3.4
2-5	27.9	3.2	7.5	5.5
6-10	37.3	5.3	8.1	5.7
≥ 11	36.9	6.3	7.2	5.4
x_{16}: Location of appointment Clinic	[<0.001]	[<0.001]	[<0.001]	[<0.001]
RMH	32.9	11.3	15.5	12.6
SMH	26.6	28.4	34.7	38.3
NDC	19.4	13.0	15.9	14.6

† % with missing data on the independent variable (x) in the analytic model
‡ Two-tail P-value from Chi-square test comparing the percentage of patients with missing data across the groupings of the candidate predictor variables for the imputation model.

Table 3: Spearman rank correlations coefficients between variables in the analytic model (columns) and candidate predictor variables (rows)†.

Candidate predictor variables for the imputation model	**x_1:Pre-consult stage of change**	**x_2:Average CPD past 6 months**	**x_3:Longest duration without smoking**	**x_4:Fagerström Score**
x_5: Gender (1=Male, 0=Female)	0.06 1069 57	**0.17** 1318 70	0.01 1251 67	**0.06** 1279 68
x_6: Year of NDC consult	0.02 1069 57	0.01 1318 70	**0.09** 1251 67	0.0001 1279 68
x_7: Age group (years)	-0.03 1069 57	-0.02 1318 70	-0.01 1251 67	-0.05 1279 68
x_8: Race: (Caucasian=1, Non-caucasian =0)	-0.04 982 52	0.03 1262 67	0.02 1207 64	-0.04 1238 66
x_9: Martial status: Single (yes=1, no=0)	0.01 1067 57	**0.06** 1318 70	0.01 1251 67	0.03 1279 68
x_{9A}: Marital status: Married or living together (yes=1, no=0)	-0.03 1067 57	-0.01 1318 70	-0.03 1251 67	-0.03 1279 68
x_{9B}: Marital status: Divorce (yes=1, no=0)	0.03 1067 57	-0.04 1318 70	0.02 1251 67	0.003 1279 68
x_{10}: Highest level of education	0.05 890 47	-0.03 1229 65	-0.0001 1183 63	-0.05 1220 65
x_{11}: Post-consultation stage of change 1=Precontemplation, 2=Contemplation 3=Preparation/Action	**0.60** 1064 57	-0.05 936 50	0.02 878 47	-0.02 891 47
x_{12}: Average CPD currently 2= 1-20, 3= 21-39, 4= ≥40	-0.05 965 51	**0.85** 1311 70	**-0.08** 1235 66	**0.61** 1267 68
x_{13}: Average CPD when smoking heaviest 2= 1-20, 3= 21-39, 4= ≥40	-0.02 947 50	**0.63** 1295 69	**-0.06** 1224 65	**0.40** 1256 67
x_{14}: Longest duration without using tobacco 1= <1 day, 2=1-30 days, 3= >1	0.01 914 49	**-0.06** 1252 67	**0.49** 1215 65	**-0.12** 1242 66
x_{15}: Number of serious stop attempts 0=none, 1= 1, 2= 2-5, 3= 6-10, 4 = ≥10	**0.1** 889 47	-0.01 1212 65	**0.12** 1178 63	-0.03 1191 63

Candidate predictor variables for the imputation model	**x_1:Pre-consult stage of change**	**x_2:Average CPD past 6 months**	**x_3:Longest duration without smoking**	**x_4:Fagerström Score**
x_{16A}: Location: Clinic (yes=1, no=0)	**-0.30** 1069 57	**-0.07** 1294 69	-0.02 1231 66	**-0.07** 1255 67
x_{16B}: Location: Rochester Methodist or St Mary's Hospital (yes=1, no=0)	**0.26** 1069 57	**0.09** 1294 69	-0.02 1231 66	**0.06** 1255 67
x_{16C}: Location: Nicotine Dependence Center (yes=1, no=0)	**0.13** 1069 57	0.01 1294 69	0.04 1231 66	0.03 1255 67

† Shown are Spearman rank correlation coefficients, number of observations used in correlation calculation, and the percentage of observations used in correlation calculation. Bolded correlation coefficients indicate those with an associated two-tail P-value <0.05 testing correlation coefficient against 0.

Table 4. Efficiency expressed as a percent for multiple imputations (MI) with k data sets and λ percent missing information. Rubin [2]

Number of datasets	λ percent of missing information				
k	10%	30%	50%	70%	90%
3	97	91	86	81	77
5	98	94	91	88	85
10	99	97	95	93	92
20	100	99	98	97	96

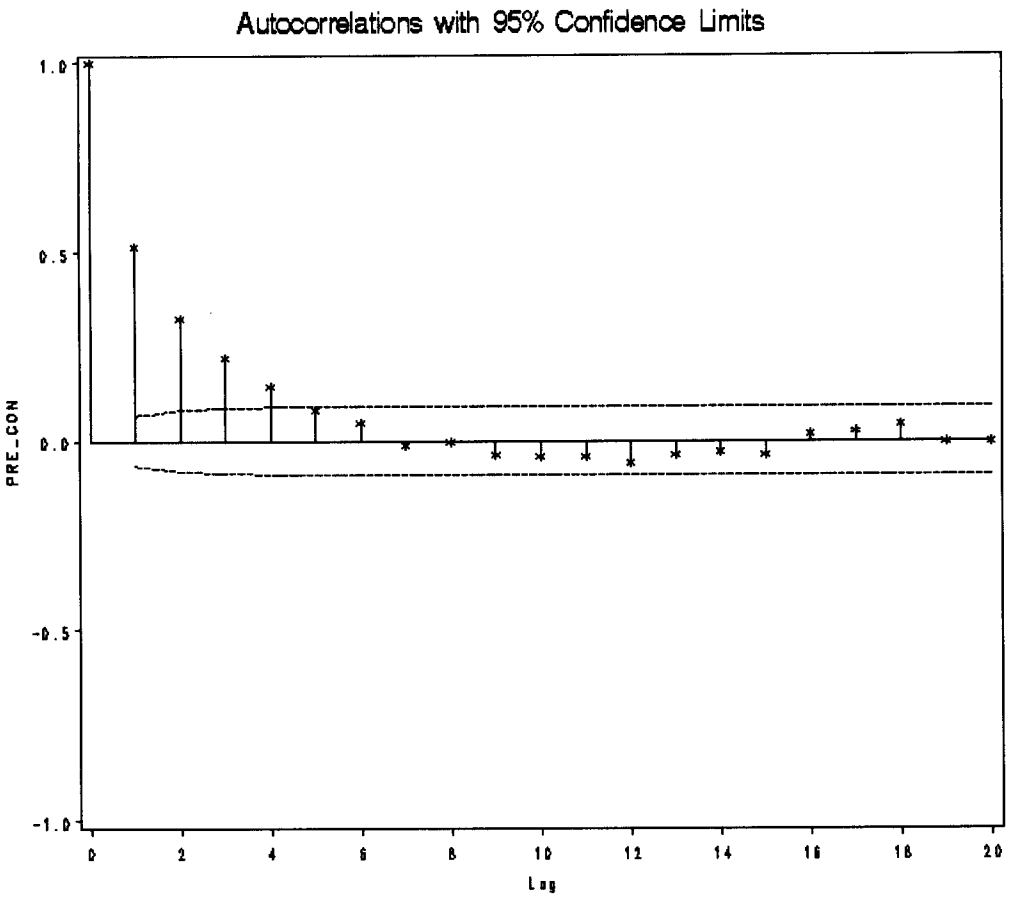

Figure 1. Autocorrelation Function (ACF) plot for x_1:pre-consult stage of change (PRE_CON) variable with 21 iterations shown. The ACF plot decays after 5 iterations, 95% CI intervals are shown with dotted lines.

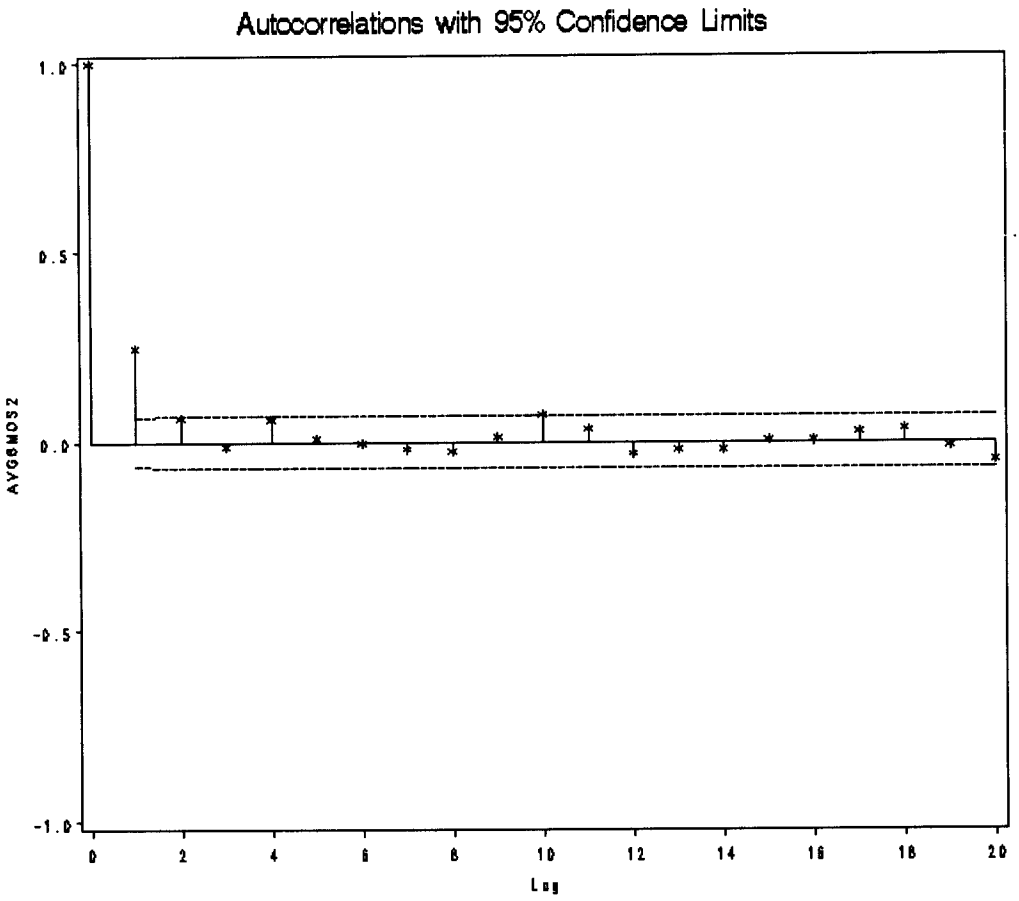

Figure 2. ACF plot for x_2:average CPD past 6 months (AVGMOS2) variable. The ACF decays after the second iteration, 95% CI intervals are shown with dotted lines.

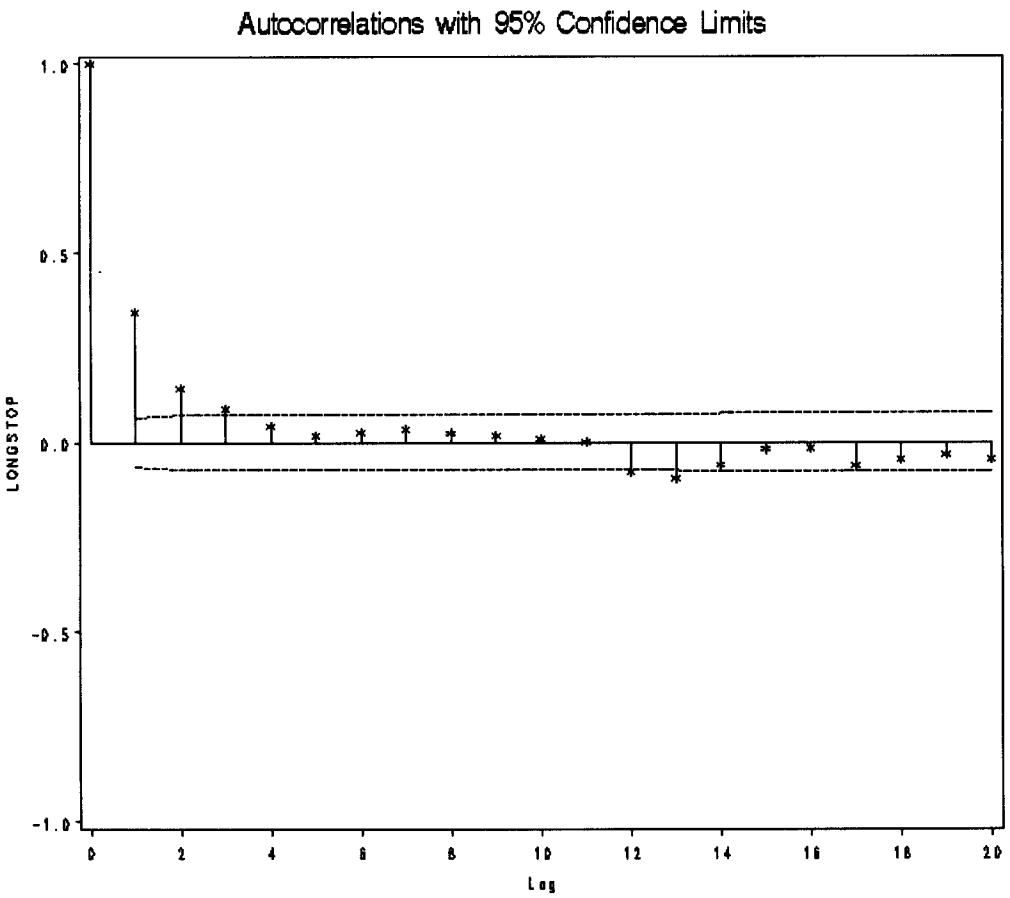

Figure 3. ACF plot for x_3: Longest duration without smoking (LONGSTOP) variable. The ACF decays after the fourth iteration, 95% CI intervals are shown with dotted lines.

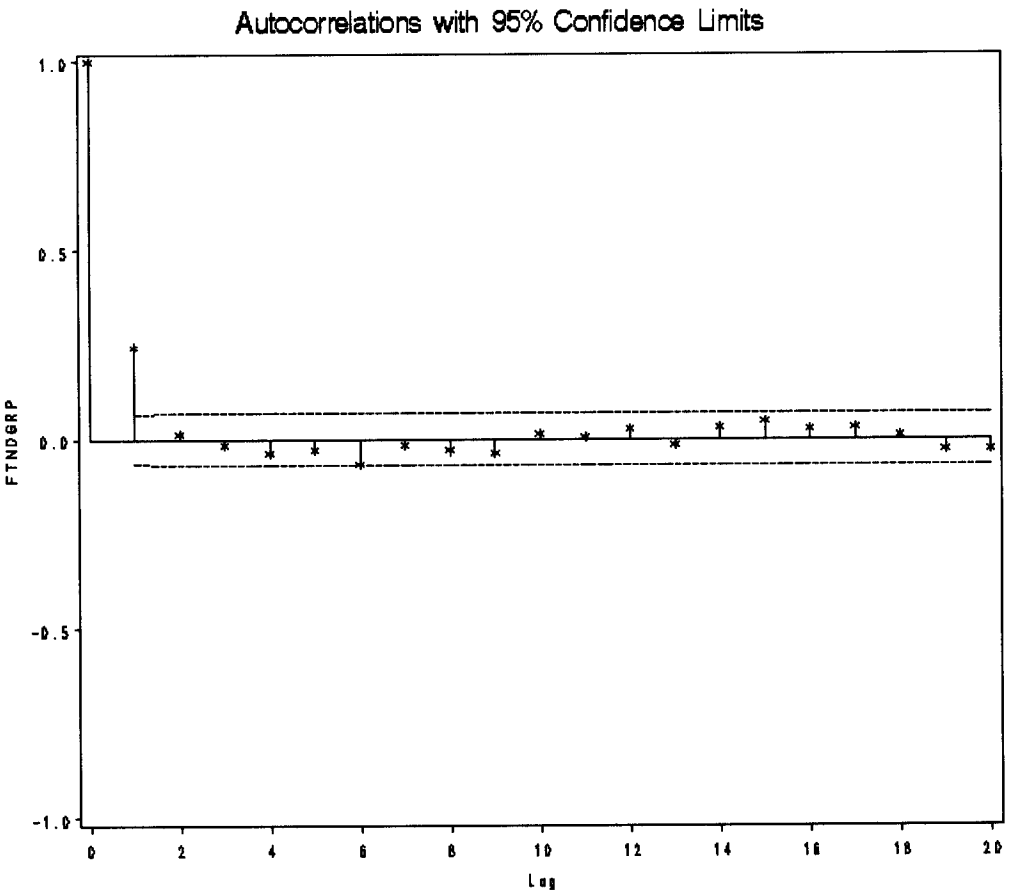

Figure 4. ACF plot for x_4: Fagerström score (FTNDGRP) variable. The ACF decays after the second iteration, 95% CI intervals are shown with dotted lines.