PharmaSUG2011 - Paper SP04

Implementation of Pattern-Mixture Models Using Standard SAS/STAT Procedures

Bohdana Ratitch, Quintiles, Montreal, Quebec, Canada
Michael O'Kelly, Quintiles, Dublin, Ireland

ABSTRACT

Methods for dealing with missing data in clinical trials have been receiving increasing attention from the regulators, practitioners and academicians in the pharmaceutical industry over the past years. New guidelines and recommendations place a great emphasis not only on the importance of carefully selecting primary analysis methods based on clearly formulated assumptions regarding the missingness mechanism, but also on the necessity to perform a range of sensitivity analyses that stress-test the results of the primary analysis under different sets of assumptions. There are many methods that could be employed for sensitivity analyses, but some of them have not yet gained a wide-spread usage, partly because of the complex underlying theory, and partly because of lack of relatively easy approaches to their implementation. In this paper, we present a new way of using standard SAS/STAT ® procedures for multiple imputation (MI) in order to implement a class of methods based on Pattern-Mixture Models (PMMs). PMMs provide a general and flexible framework for sensitivity analyses that allows formulating assumptions regarding missing data in a transparent and clinically interpretable manner. Our implementation strategy is based on the core functionality available in PROC MI and MIANALYZE procedures and does not require any additional statistical programming outside DATA steps and sorting. A specific PMM-based method that we discuss here relies on clear and realistic clinical assumptions, while the general principles of our approach can be used to implement a range of other methods with different sets of assumptions.

INTRODUCTION

Some amount of missing data is unavoidable in virtually all clinical trials, no matter how well designed. Methodological research for missing data has been active for decades, while practical application of some advanced methods has been slow in the pharmaceutical industry. Recently, this topic received a great deal of renewed attention from regulatory authorities which resulted in new guidelines from European Medicines Agency [1] and an FDA-mandated panel report from National Research Council in US [2]. Both documents express a requirement for statistical analyses of clinical trial data that can be justified as assuring the validity of trial conclusions in the presence of missing information.

Missing data complicate the statistical analysis of clinical trials because the analysis methods must make certain assumptions about the missingness mechanism and unobserved values to deal with incomplete cases. Usually, a primary analysis will be based on a missing at random (MAR) assumption, which means that missingness is independent of the unobserved outcomes after accounting for the appropriate observed data in the model. Historically, primary analyses often made an even stronger assumption of missing completely at random (MCAR), which assumes that missingness does not depend on either observed or unobserved data. These assumptions are fundamentally unverifiable by virtue of our inability to collect the data either to support or refute them. Therefore, new guidelines place a great emphasis on the necessity to perform sensitivity analyses under different assumptions regarding missing data in order to establish a range of conditions under which study conclusions are robust as well as conditions under which these conclusions can be overturned. Regulatory recommendations stress the importance of formulating the assumptions for various sensitivity analyses in such a way that they could be easily interpreted from a clinical point of view, and thus their clinical plausibility could be discussed.

Sensitivity analyses are often expected to address the possibility of the data being missing not at random (MNAR). MNAR means that missingness depends on the unobserved values, and cannot be predicted solely based on subject's observed data. Several types of statistical models have been proposed to analyze clinical study data under such assumptions. Most prominent are selection models, shared parameter models and pattern-mixture models. Each of them has some advantages and disadvantages. This paper is concerned with PMMs, their application to randomized, parallel-arm, longitudinal clinical studies, and their implementation using standard SAS/STAT procedures using multiple imputation methodology. One of the advantages of PMMs is that they allow transparent and clinically interpretable formulations of the assumptions regarding unobserved data. The importance of communicating the assumptions underlying statistical analyses in terms that can be understood by clinicians has been emphasized by regulatory authorities and various industry working groups, e.g., [1, 2].

In most randomized clinical trials that collect longitudinal data (at several study visits), the great majority of missing data are caused by subjects discontinuing from the trial prior to a study visit at which primary endpoint data are collected. The resulting missing data will have a monotone pattern, meaning that, once a subject has a missing data for some visit, data will be missing for all subsequent visits. This paper is primarily concerned with dealing with this type of missing data patterns. Typically, there also is some small amount of non-monotone missing data (when subjects skip intermediate visits but return for evaluations at subsequent visits). We briefly discuss how to deal with such missing data as well.

In this paper, we present the details for implementation of one method for sensitivity analysis within the PMM framework. This method assumes that after discontinuation, subjects discontinued from the experimental treatment arm will exhibit an evolution of the disease similar to subjects in the control arm. An approach to the implementation of the PMM-based methods that we discuss here is based on some general principles proposed in the literature in the past [3, 4]. In this paper we suggest what may be a novel way of using SAS functionality for multiple imputation, namely methods for imputing monotone missing data patterns available in PROC MI. Using multiple imputation as integral part of a PMM-based analysis has a great advantage in that MI procedures automatically take care of producing correct variance estimates that account for uncertainty associated with the imputation process. The approach we propose in this paper uses only core SAS/STAT functionality (readily available in PROC MI and MIANALYZE) without the need for any additional statistical programming using, for example, IML. The key feature of our implementation strategy is to perform imputation using a sequence of calls to PROC MI with a MONOTONE statement, and not a single call as is typically done to impute all values at once. The sequential procedure we use preserves the underlying principles of sequential (chain) imputation for monotone missing patterns that is automatically implemented as an option in PROC MI. At the same time, breaking out the imputation process into multiple procedure calls provides for a greater flexibility in stress-testing a primary MAR or MCAR analysis by imposing suitable MNAR assumptions to evaluate the robustness of the study findings. The MONOTONE statement in PROC MI allows for the imputation of both continuous and categorical variables. The logic of the implementation strategy that we propose in this paper would be the same for both continuous and categorical variables. It would allow the application of PMMs to categorical outcomes without having to compromise or approximate the expected distribution for these variables. Such approximation is necessary for example when the MCMC option is used in PROC MI.

OVERVIEW OF PATTERN-MIXTURE MODELS

PMMs represent a general framework that accommodates several distinct variants and allows several approaches to their implementation. They all share the following basic principle. Let the entire data matrix Y be separated into two components: Y_{obs}, representing observed data and Y_{mis}, representing missing (unobserved) data. Let X be a set of (observed) covariates. Let R represent a matrix of indicators of missingness. PMMs decompose the joint probability of data and missingness as follows:

$$p(Y_{obs}, Y_{mis}, R|X) = p(R|X)\ \ p(Y_{obs}, Y_{mis}|\ R, X) \qquad (1)$$

In the PMM framework, subjects are grouped into cohorts so that subjects in the same cohort share the same pattern of missing data. Patterns may be based on different characteristics, for example, time of discontinuation from the clinical study, reason for discontinuation, or even treatment arm to which subjects were randomized. The probability distribution p(R|X) can be viewed as a probability distribution of various missingness patterns. A complete data analysis model $p(Y_{obs}, Y_{mis}|\ R, X)$ is then estimated within each pattern (cohort). The pattern-specific estimates are not typically of interest, so the average estimates across the missing data patterns are obtained to yield an overall result. Averaging is accomplished by the weighting factor p(R|X).

Pattern-mixture models are, by definition, under-identified because patterns with missing data typically have some parameters of the $p(Y_{obs}, Y_{mis}|\ R, X)$ model that cannot be estimated from data due to incomplete data within that pattern. However, this should not be regarded as a disadvantage or impediment for their practical application. On the contrary, in order to estimate PMMs, one simply needs to explicitly impose some assumptions regarding the inestimable parameters, and this explicitness helps to clearly interpret the basis and results of the analyses. Such assumptions are referred to in the PMM literature as "identifying restrictions". There are several methodologies for imposing these identifying restrictions. We focus on one of them that essentially links missing data to observed data across different patterns. Probability distribution in equation (1) can be further decomposed as follows:

$$p(Y_{obs}, Y_{mis}, R|X) = p(R|X)\ \ p(Y_{obs}, Y_{mis}|\ R, X) = p(R|X)\ \ p(Y_{obs}\,|\,R, X)\ \ \ p(Y_{mis}\,|\,Y_{obs}, R, X) \qquad (2)$$

In this formulation, $p(Y_{obs}\,|\,R, X)$ represents a model for available data within each pattern, and $p(Y_{mis}\,|\,Y_{obs}, R, X)$ represents a model for missing data conditioned on observed data within each pattern. A link can be defined between the two, so that the model $p(Y_{obs}\,|\,R, X)$ from one pattern could be used to identify $p(Y_{mis}\,|\,Y_{obs}, R, X)$ (or impute values) in another pattern. The following approach, described in [4], Section 16.5, can be used to implement this strategy in practice as follows.

i. Choose missing data patterns, i.e., cohorts of subjects that share similar missingness characteristics. Note that subjects that completed a study constitute a "completers" pattern, which may be further subdivided by treatment arm.

ii. For each pattern, specify a link between the distribution of unobserved data and the distribution of data in patterns where the corresponding data items are observed. In the equation (2), it corresponds to specifying a form of an imputation model $p(Y_{mis}\,|\,Y_{obs}, R, X)$ for each pattern and its relationship to models in other patterns.

iii. Estimate one or more models based on the observed data.

iv. According to the link function in (ii), use standard multiple imputation techniques to impute missing data in each pattern with missing data based on draws from model(s) estimated in (iii).

v. Analyze multiply-imputed datasets by a method of choice for complete data and combine the results based on a standard MI methodology.

Based on this general framework, different sensitivity analyses can be performed by making different choices for pattern definition in (i) and link functions in (ii) above. Patterns and link functions provide the means of embedding various clinical assumptions into the statistical model. Below we will describe in detail one variant of the approach described above and will mention briefly other variants where the assumptions regarding unobserved data have clear clinical interpretations.

EXAMPLE DATASET

Implementation of a PMM-based analysis will be illustrated using an example study where efficacy endpoint is stored in a SAS dataset, DATAIN, with the following variables:

- SUBJID – subject ID number
- TRT - treatment arm (with 0 representing control (placebo) treatment arm, and 1 representing an active experimental treatment arm)
- SCORE_0, SCORE_1, SCORE_2, SCORE_3 – efficacy scores at time-points 0 (baseline), and post-baseline visits 1, 2, and 3. These are continuous variables, where higher scores represent more favorable outcomes.
- LASTVIS – last study visit attended by a subject (takes values 0, 1, 2, or 3)

In our dataset, all subjects have non-missing values at baseline (SCORE_0). One group of subjects received study treatment, but discontinued from the study prior to the first scheduled post-baseline visit. These subjects have missing values of SCORE_1, SCORE_2, and SCORE_3. Other subjects discontinued from the study either after time-point 1 or 2 and have missing values of SCORE_2 and/or SCORE_3. A summary for the percentages of subjects who discontinued and subjects who completed the study by time-point and treatment arm is provided in Table 1.

Table 1: Percentage of subjects discontinuing from the study and study completers

	Discontinued Subjects (Cumulative)	
Time-point	Placebo Arm	Active Treatment Arm
0 (Baseline)	0	0
1	13%	8%
2	19%	14%
3	21%	20%
	Study Completers	
	79%	80%

In the next section, we will briefly summarize multiple imputation functionality available in SAS and illustrate its usage on this example dataset.

STANDARD MULTIPLE IMPUTATION IN SAS

Multiple imputation can be performed in SAS [5] using a general three-step approach. First, PROC MI is applied to an input dataset containing some missing values, which results in creation of multiple copies of this dataset. All copies contain identical values of the non-missing data items, but different values imputed for missing items. Each of these imputed datasets can subsequently be analyzed using standard SAS procedures, e.g., PROC GLM, PROC MIXED, PROC LOGISTIC, etc. Results from all imputed datasets are then combined together for overall inference using PROC MIANALYZE.

There are several methods available for imputation in PROC MI. Which methods can be used depends on the type of variables that need to be imputed (continuous or categorical) and on the pattern of missing data. Note that there is no direct correspondence between the definition of patterns in PMMs and the way this term is used in PROC MI. In PMMs, pattern represents a group of subjects that share some general common characteristics related to drop-out. In PROC MI, pattern refers to a configuration or position of missing values when analysis variables are arranged in a certain order.

PROC MI distinguishes two types of missingness patterns: non-monotone and monotone. For example, in the context of clinical trial data, let us assume that an input dataset has a horizontal structure so that there is one record per subject, and assessments at different study visits are stored in separate variables with the order of variables corresponding to the chronological order of study visits. A dataset is said to have a monotone missing pattern if missing values always occur at the end of data records. In the case of our example, it would correspond to subjects that prematurely discontinued from the study and have a number of visits with missing data after their discontinuation. For a dataset to have a monotone missing pattern, all subjects with some missing data would need to exhibit this configuration of missingness. On the other hand, a dataset is said

to have a non-monotone missing pattern if some subjects have missing values for intermediate visits, but have available data at subsequent visits.

If a dataset has a non-monotone pattern of missingness, the only available imputation method in PROC MI is based on the Monte Carlo Markov Chain (MCMC) methodology and assumes a multivariate normal distribution over all variables included in the imputation model. For datasets with monotone missing patterns, a number of different methods are available. There are several choices for categorical and continuous variables, including some parametric and non-parametric methods. The MCMC method can also be applied to datasets with monotone missing patterns.

With the MCMC method, values for all missing variables are sampled simultaneously from the multivariate normal distribution. In contrast, methods for monotone missingness implemented in PROC MI are based on a sequential procedure where variables are imputed one at a time using univariate imputation models. Consider a simple example of a call to PROC MI with a MONOTONE statement.

```
proc mi data=DATAIN out=DATAOUT;
    var TRT SCORE_0 SCORE_1 SCORE_2;
    monotone regression;
run;
```

Variables SCORE_1 and SCORE_2 contain missing values for some subjects, which PROC MI will impute using linear regression models as requested in a MONOTONE statement. Based on the order of variables specified in the VAR statement, PROC MI imputes variables from left to right. First, variable SCORE_1 will be imputed using a regression model that includes variables preceding it in the VAR statement as predictors (TRT and SCORE_0). Then, variable SCORE_2 will be imputed using a regression model with TRT, SCORE_0, and SCORE_1 as predictors. For subjects that had values of both SCORE_1 and SCORE_2 missing, values of SCORE_1 imputed on the previous step are used to predict the values of SCORE_2 from this model.

This sequential imputation approach is considered to perform well with monotone missingness [4, 8, 10]. For non-monotone missing data, variants of a sequential (chain) procedure have also been suggested, e.g., [9]., but their theoretical properties and empirical performance are not yet well understood, and they are not currently implemented in PROC MI.

We will now illustrate how a standard multiple imputation methodology could be implemented using SAS procedures on our example dataset introduced in the previous section. Our dataset contains one subject that missed an intermediate visit and has a missing value of SCORE_2, but a non-missing value of SCORE_3. With this subject, the dataset has a non-monotone pattern of missingness in PROC MI terminology. In order to obtain a monotone missing data pattern, we can first use the MCMC method in PROC MI to partially impute data as illustrated below (note the usage of `impute=monotone` option in combination with the MCMC statement).

```
proc mi data=DATAIN out=DATAIN_MONO nimpute=100 seed=123;
    var TRT SCORE_0 SCORE_1 SCORE_2 SCORE_3;
    mcmc chain=multiple impute=monotone;
run;
```

As a result of this procedure call, only the missing value of SCORE_2 for a single subject with a missing intermediate visit will be imputed. The resulting output dataset DATAIN_MONO will have 100 copies of the original dataset where each copy will have a monotone missing pattern.

Assumptions underlying this partial imputation step are such that subjects with missing data follow the same model as other subjects in their respective treatment arm that have complete data. This is a standard way of using multiple imputation under the missing at random (MAR) assumption, meaning that missingness is independent of the unobserved values after accounting for the observed values in the imputation model. It can be argued that in most situations, this is a reasonable assumption for this partial imputation process because subjects tend to miss intermediate visits due to scheduling conflicts or other reasons unrelated to their medical condition under study. We note that there may be studies where intermittent missing values have a particular interpretation that may not be well served by MCMC imputation. However, in general, the fact that these subjects continue participating in the study and return for subsequent visits can justify modeling these intermediate missing values based on a general MAR-based imputation model including treatment arm effect, even if the rest of the missing data (monotone patterns) will be imputed based on different assumptions. The proportion of non-monotone missing data is typically very small compared to the number of subjects that discontinue from the study permanently. Because of these reasons, a small amount of non-monotone missingness will typically have a small effect on the results of analysis at the final time-point and imputing these few values using the MCMC method above should not compromise the validity of the sensitivity analyses.

The rest of the missing data in our example can now be imputed using a method for monotone missingness. Using a standard MAR-based multiple imputation approach, it can be done with the following call to PROC MI where the output dataset, DATAIN_MONO, from the previous call using MCMC for partial imputation serves as an input dataset. This dataset already contains multiple (100) partially imputed datasets, so the subsequent call to PROC MI to complete the imputation will use a BY _Imputation_ statement and request one imputed dataset (`nimpute=1` option) within each BY group.

```
proc sort data=DATAIN_MONO; by _Imputation_ TRT; run;
```

```
proc mi data=DATAIN_MONO out=DATAIN_REG seed=465 nimpute=1;
    by _Imputation_;
    var TRT SCORE_0 SCORE_1 SCORE_2 SCORE_3;
    class TRT;

    monotone regression;
run;
```

The output dataset DATAIN_REG contains 100 fully imputed datasets ready to be analyzed by standard SAS procedures. The following steps demonstrate how to conduct an ANCOVA analysis based on these multiply-imputed data and how an overall result can be obtained using PROC MIANALYZE.

First, let us transform the imputed dataset so that for each subject, each post-baseline time-point is represented by a separate record, and ANCOVA analysis could be performed for each time-point using a BY statement. On each record, we will compute a change from baseline in the efficacy score.

```
data DATAIN_REG1; set DATAIN_REG;
       TIMEPTN=1; SCORE_C = SCORE_1 - SCORE_0; OUTPUT;
       TIMEPTN=2; SCORE_C = SCORE_2 - SCORE_0; OUTPUT;
       TIMEPTN=3; SCORE_C = SCORE_3 - SCORE_0; OUTPUT;
run;
```

ANCOVA analysis for change from baseline in the efficacy score at each time-point is performed using PROC MIXED below and is based on a model including treatment arm as a fixed effect and a baseline efficacy score as covariate. LSMEANS statement is used to request least squares estimates for means of the change from baseline in each treatment arm as well as the mean difference between the changes from baseline in the active treatment arm versus placebo. These estimates are captured using ODS output datasets. A BY statement in the procedure below includes _Imputation_ variable, therefore the analysis is performed within each of the 100 imputed datasets.

```
proc sort data=DATAIN_REG1; by _Imputation_ TIMEPTN TRT; run;

proc mixed data=DATAIN_REG1;
       by _Imputation_ TIMEPTN;
       class TRT;
       model SCORE_C = TRT SCORE_0 / solution covb;
       lsmeans TRT / diff=control('0') cl;
       ods output Diffs=DIFF_MI LSMeans=LSM_MI;
run;
```

Results of the ANCOVA analysis on 100 imputed datasets can now be combined to derive an overall result. This is done by applying PROC MIANALYZE to the ODS output datasets DIFF_MI and LSM_MI produced by PROC MIXED above.

```
proc sort data=DIFF_MI; by TIMEPTN _Imputation_; run;

proc mianalyze parms(classvar=full)=DIFF_MI;
       class TRT ;
       modeleffects TRT;
       ods output ParameterEstimates=DIFF_MIAN;
       by TIMEPTN;
run;

proc sort data=LSM_MI; by TIMEPTN _Imputation_; run;

proc mianalyze parms(classvar=full)=LSM_MI;
       class Trt ;
       modeleffects TRT;
       ods output ParameterEstimates=LSM_MIAN;
       by TIMEPTN;
run;
```

The results of this analysis are summarized in Table 2. In the next section, we will discuss how to implement a pattern-mixture model with a control-based pattern imputation using PROC MI functionality. The implementation of PMMs described below relies on the sequential imputation methodology implemented in PROC MI for datasets with monotone missingness. The idea of using sequential procedures has been suggested in the context of PMMs by several authors (see, e.g., [2, 3, 4]).

PATTERN-MIXTURE MODEL WITH CONTROL-BASED PATTERN IMPUTATION

The main idea of this method was introduced by Little and Yau (1996) [4]. The original method suggests using sequential regression and multiple imputation methodology to impute missing values after subject's discontinuation from the trial based on "as treated" model, using actual dose after drop-out if it is known, or based on some plausible assumptions if unknown. In most clinical trials, subjects stop taking experimental medication after discontinuation. In this case, the "as treated" model for discontinued subjects would be based on the idea that subjects are taking a zero dose of the experimental treatment. Therefore, it might often be reasonable to assume that after withdrawal from the study, subjects from the experimental treatment arm (no longer receiving the experimental treatment) will exhibit the same future evolution of the disease as subjects on the control treatment (who are also not exposed to the experimental treatment). Subjects that discontinue from the control arm are assumed to evolve in the same way as control subjects that remain in the study. This assumption is particularly attractive for studies where there is a control treatment that consists exclusively of placebo. Another example where such strategy would be suitable is a study with a standard-of-care control treatment and where subjects discontinued from the experimental arm switch (or may be assumed to switch) to the standard-of-care treatment. Such assumptions would tend to move the estimate of the treatment effect (difference between the experimental and control treatments) towards a smaller value compared to the analyses based on MAR-based methods, such as mixed models with repeated measures (MMRM) or standard MI models. These latter models essentially assume that subjects from the experimental treatment arm continue to reap benefits from the experimental treatment after discontinuation.

Note that if subjects who discontinue are expected to change to a treatment other than the control treatment, as can be the case, for example, in oncology trials, then the assumption that withdrawals follow the pattern of the control group may not be plausible clinically.

The approach proposed by Little and Yau has been recently explored by Roger et al. [6, 7] who suggested two ways of implementing this idea: one involving multiple imputation and another based on MMRM analyses.

In an MI-based implementation, Roger et al. suggested building a model for a joint distribution of the outcome at all time-points based on the available data from control subjects and use multiple imputation to impute missing values of all discontinued subjects (from both the control and experimental treatment arms) using this model. Roger developed a SAS macro to implement this approach using PROC MI's MCMC statement, where MCMC's estimates of the means and covariance matrix of the multivariate normal distribution estimated from control subjects is captured in an output dataset, and then is used to impute data for subjects from the experimental treatment arm using the IML functionality.

In an MMRM-based implementation, Roger suggested to perform a maximum likelihood analysis (without explicit imputations) and to isolate a contrast in which a parameter estimate for the experimental treatment effect is replaced by a linear combination of the regression coefficients: a portion of a parameter estimate corresponding to the experimental treatment effect is used for experimental arm completers, and a portion of a parameter corresponding to the control treatment effect is used for experimental arm drop-outs. The "portions" of each parameter estimate in this linear combination are determined by the proportion of completers and withdrawals in the experimental treatment arm.

In this paper, we present another method of implementing the same general idea, but using PROC MI's methodology for imputation of monotone missing data patterns (available with the MONOTONE statement) to impute the outcome variable at consecutive visits in a sequential (chain) manner. This method has an advantage over Roger's implementation based on the MCMC method in that the method described here allows modeling and imputation of categorical response variables using methods appropriate for categorical variables instead of approximations based on multivariate normal distribution. Availability of a CLASS statement with the MONOTONE statement in PROC MI (in contrast to the MCMC statement that does not allow CLASS variables) also provides for an easier modeling of the categorical predictor variables. Another advantage of our approach is that it does not require any additional programming with IML.

In control-based pattern imputation, our intention is to make no direct use of observed data from the experimental treatment arm for estimating the imputation model. We achieve this by calling PROC MI in such a way that it builds its imputation model only on data from the control arm, while it imputes missing data in both control and experimental treatment arms using a single control-based imputation model. This can be achieved with a sequence of calls to PROC MI as will be described below.

In contrast, the standard use of PROC MI with the MONOTONE statement would be based on a single call to PROC MI to impute all missing values (all subjects and all time-points) as illustrated in the previous section. In this case, treatment arm was included as an effect in the imputation model, and thus subjects from each treatment arm followed an "arm-specific" model. Note that if we simply excluded the treatment arm effect in this single-call application of PROC MI, the imputation model would have been estimated based on available data from all subjects included in the input dataset (from both control and experimental arms) albeit not modeling the correlation between treatment and efficacy scores.

In order to implement the control-based pattern imputation, we will break the imputation process into a sequence of multiple calls to PROC MI, where each call is intended to impute missing values at one time-point only. The general logic of such a strategy is as follows.

i. With each call to PROC MI, impute only one time-point. That is, include only one variable corresponding to the time-point that needs to be imputed in the VAR statement (plus predictors), while respecting the order in which PROC MI would have done it in a single call (all time-points in chronological order).

ii. When imputing missing values for time-point *t*, the input dataset should include all control subjects, but only those subjects from the experimental arm that have values at time-point *t* missing (only those that need imputation at time-point *t*). Since subjects from the experimental arm with non-missing values at time-point *t* are not included in the input dataset, they will not contribute to the estimation of an imputation model for time-point *t*. Imputation model will be estimated using control subjects only, while this call to PROC MI will impute missing data at time-point *t* for all subjects who need imputation at that time-point. This way, subjects from experimental arm will be imputed based on the control subjects' model. Note that treatment arm should not be included as an effect in this model.

iii. Repeat (ii) for all other time-points sequentially. Subjects whose missing values were imputed in the last call to PROC MI will be included in the input dataset for the next call to PROC MI. Thus data for time-point *t,* filled in during the last call, will be used for predictor variables in the next call to PROC MI (for time-point *t+1*), which is consistent with the internal workings of a single call to PROC MI to impute all time-points automatically.

This strategy is illustrated below using an example dataset introduced in the previous section. Assume that non-monotone missing data have already been imputed using the MCMC method and the result of this partial imputation is stored in the dataset DATAIN_MONO, which contains a variable _Imputation_ to distinguish between multiple copies of the original input data. We can now use a monotone imputation method to impute the rest of the missing data. First, we will impute missing data at time-point 1 (the first time-point that has some missing data). In order to do it using the control-based imputation method, we will separate our data into two datasets: DATAIN_MONO_IMP1, containing all control subjects and those subjects from the experimental arm that have values at time-point 1 missing; and DATAIN_MONO_REST1, containing the rest of the subjects from the experimental arm (those with non-missing SCORE_1).

```
data DATAIN_MONO_IMP1 DATAIN_MONO_REST1;
    set DATAIN_MONO;
    if TRT = 1 and LASTVIS >=1 then output DATAIN_MONO_REST1;
    else output DATAIN_MONO_IMP1;
run;
```

We will now call PROC MI to impute missing data at time-point 1 based on the model estimated exclusively from control subjects with non-missing values at time-point 1.

```
proc mi data=DATAIN_MONO_IMP1 out=DATAIN_REG_IMP1 nimpute=1 seed=234;
    by _Imputation_;
    var SCORE_0 SCORE_1;
    monotone reg(SCORE_1);
run;
```

The syntax `monotone reg(SCORE_1)` requests a default regression model for imputing SCORE_1 – a model that includes all variables preceding SCORE_1 in the VAR statement as effects (SCORE_0 in this case).

The following data step assembles back a dataset containing all subjects.

```
data DATAIN_IMP1;
    set DATAIN_MONO_REST1 DATAIN_REG_IMP1;
run;
```

Now we will proceed with imputing missing values for time-point 2. Dataset DATAIN_IMP1 will be used for input to the next PROC MI call, but first, we will separate it into two datasets: DATAIN_MONO_IMP2, containing all control subjects and those subjects from the experimental arm that have values at time-point 2 missing; and DATAIN_MONO_REST2, containing the rest of the subjects from the experimental arm.

```
data DATAIN_MONO_IMP2 DATAIN_MONO_REST2; set DATAIN_IMP1;
    if TRT = 1 and LASTVIS >= 2 then output DATAIN_MONO_REST2;
    else output DATAIN_MONO_IMP2;
run;
```

We will call PROC MI to impute missing data at time-point 2 based on the model estimated exclusively from control subjects with non-missing data at time-point 2. A `monotone reg(SCORE_2)` statement below will result in using an imputation model for SCORE_2 with SCORE_0 and SCORE_1 as predictors.

```
proc sort data= DATAIN_MONO_IMP2; by _Imputation_; run;

proc mi data=DATAIN_MONO_IMP2 out=DATAIN_REG_IMP2 nimpute=1 seed=345;
    by _Imputation_;
    var SCORE_0 SCORE_1 SCORE_2;
    monotone reg(SCORE_2);
run;
```

As previously done, we can assemble back a dataset containing all subjects:

```
data DATAIN_IMP2;
    set DATAIN_MONO_REST2 DATAIN_REG_IMP2;
run;
```

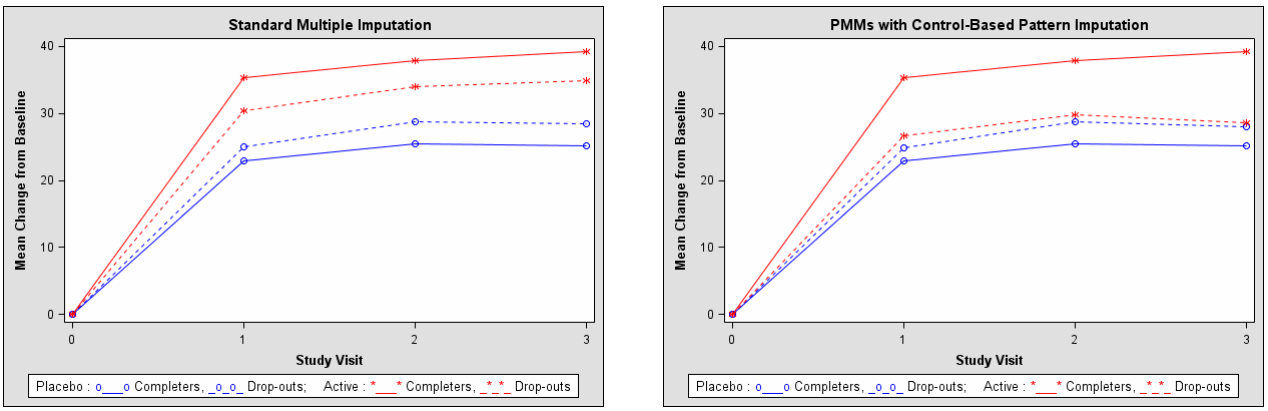
The same procedure should be applied in order to impute missing values for the last time-point 3:

```
data DATAIN_MONO_IMP3 DATAIN_MONO_REST3; set DATAIN_IMP2;
    if TRT = 1 and LASTVIS = 3 then output DATAIN_MONO_REST3;
    else output DATAIN_MONO_IMP3;
run;


proc sort data= DATAIN_MONO_IMP3; by _Imputation_; run;

proc mi data= DATAIN_MONO_IMP3 out= DATAIN_REG_IMP3 nimpute=1 seed=456;
    by _Imputation_;
    var SCORE_0 SCORE_1 SCORE_2 SCORE_3;
    monotone reg(SCORE_3);
run;

data DATAIN_IMP3;
    set DATAIN_MONO_REST3 DATAIN_REG_IMP3;
run;
```

If the study contained more time-points, the above procedure would be repeated until all time-points are imputed. For this example, dataset DATAIN_IMP3 is the final dataset containing 100 imputed datasets where all missing values are filled in and ready for analysis.

Figure 1 illustrates how imputations from the control-based pattern imputation (right panel) differ from those produced by a standard multiple imputation approach (left panel). On these graphs, solid lines represent mean change from baseline in efficacy score for subjects who completed the study (observed data), and dashed lines represent means of discontinued subjects whose data were imputed. With standard multiple imputation, active treatment arm drop-outs follow a slope of completers in their treatment arm. With control-based pattern imputation, imputed values in the active treatment arm follow the trajectory of placebo subjects.

Figure 1: Mean change from baseline in efficacy score resulting from standard MI and PMMs with Control-Based Pattern Imputation.



ANCOVA analysis on these multiply-imputed data can be performed in exactly the same manner as shown in the previous section with standard MI.

The results of analysis with control-based pattern imputation and with the standard multiple imputation are summarized in Table 2. As expected, the least squares estimates of the mean difference between the active treatment and placebo are smaller for the control-based pattern imputation method compared to standard MI. Nevertheless, in this example, the differences remain statistically significant under the assumptions of this sensitivity analysis. Clinical judgment would need to be applied to decide whether the difference remains clinically significant under these assumptions.

Table 2: Results of the ANCOVA analysis using standard MI and PPM with control-based pattern imputation.

	Standard Multiple Imputation			PMM with Control-Based Pattern Imputation		
Time-Point	Least Squares Mean Difference (Active – Placebo) (95% CIs)	Standard Error	P-value	Least Squares Mean Difference (Active – Placebo) (95% CIs)	Standard Error	P-value
1	10.600 (5.082 16.119)	2.815465	0.0002	9.870 (4.156, 15.584)	2.914882	0.0007
2	10.936 (4.909, 16.962)	3.074399	0.0004	10.061 (3.961, 16.160)	3.111254	0.0012
3	12.130 (5.726, 18.535)	3.266567	0.0002	10.924 (4.537, 17.311)	3.257907	0.0008

This sensitivity analysis stress-tests the MAR assumption that withdrawals will tend to have efficacy similar to subjects who remain in the study in their respective treatment arms. Control-based pattern imputation assumes, on the contrary, that after discontinuation subjects on the experimental treatment who withdraw will tend to have efficacy close to subjects on the control treatment. It will thus tend to provide a reduced estimate of treatment effect that will be less in favor of experimental treatment compared to the standard estimate under MAR assumptions. However, it does not push the analysis into a more extreme scenario where subjects who discontinue from the experimental arm could be doing worse than subjects in the control arm. This might happen, for example, if control treatment is an active marketed drug and the experimental treatment is a combination of an active drug used in the control arm and a new experimental treatment. If the experimental treatment interferes with the effect of the control drug and reduces its efficacy in a sub-population of subjects that discontinue, then subjects from the experimental arm could do worse than control subjects, and the control-based pattern imputation would underestimate this effect. Some methods suggested in the Conclusions section would allow pushing the assumptions into this direction and could be implemented in a similar framework.

In this example, we were using a regression method for imputing variables with missing data for illustration purposes, but it could be replaced by other methods for monotone patterns available with the MONOTONE statement. Thus, the same strategy could be used to impute categorical outcome variables.

CONCLUSION

This paper presents a way of implementing sensitivity analyses for dealing with missing data in clinical trials. The method discussed here belongs to the class of pattern-mixture models and uses an implementation strategy based on an idea first described by Little and Yau. A particular variant of this strategy presented here is based on the multiple imputation methodology.

We presented the details for implementation of one specific method, control-based pattern imputation, using SAS procedures PROC MI and MIANALYZE. A variety of other methods (with different clinical assumptions) can be implemented using a similar strategy. For example, another approach that was recommended in [4] would assume that subjects that discontinue from the study are similar to other subjects that discontinued from their respective treatment arm (although at a later time), but differ from the completers. This strategy is sometimes referred to in the literature as neighboring pattern imputation. Yet another method recommended in [2, 4] would be based on an assumption that after discontinuation, subjects from the experimental treatment arm will follow a model that is to some extent similar to other subjects in their arm, but their condition will be somewhat worse than that of subjects who continue on experimental treatment. A numeric parameter, with a clear clinical interpretation, could be used to quantify this difference. A range of analyses with different (increasing) settings of this parameter can be carried out in order to assess the robustness of the study conclusions and find a so-called tipping point at which the study conclusions are overturned. As the value of this parameter increases, it is possible to model a scenario where subjects discontinued from the experimental treatment arm do worse than the control subjects. These and other methods could be easily implemented in SAS using our strategy of multiple successive calls to PROC MI, and are the subject of ongoing work by the authors.

We hope that the fact that this implementation is based on the standard functionality available in PROC MI and does not require manual statistical programming beyond some simple data-step manipulations of the input datasets will make the PMM-based analyses more accessible to the practitioners in the pharmaceutical industry. The approach that we presented here can be used to analyze both continuous and categorical outcome variables using PMM-based methods. As illustrated by several sensitivity analyses suggested above, PMMs deserve a wider usage because they allow formulating the assumptions about the missing data in a clear manner that can be easily understood by clinicians. . The multiple imputation approach to PMMs that is described here gives clinically useful “what-if” estimates of treatment effect and gives reasonable estimates of the variance of those “what-if” estimates, allowing formal stress-testing of the assumptions of the primary analysis.

REFERENCES

[1] European Medicines Agency. Guideline on Missing Data in Confirmatory Clinical Trials. July 2010.

[2] National Research Council (1010). The Prevention and Treatment of Missing Data in Clinical Trials. Panel on Handling Missing Data in Clinical Trials. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academic Press.

[3] Little, R., Yau, L. Intent-to-Treat Analysis for Longitudinal Studies with Drop-Outs. Biometrics, 1996, vol. 52, 1324-1333.

[4] Molenberghs, G., Kenward, M. G., Missing Data in Clinical Studies. Wiley, 2007.

[5] SAS Institute Inc. 2008. SAS/STAT ® 9.2 User's Guide. Cary, NC: SAS Institute Inc.

[6] Roger, J., Ritchie, S., Donovan, C. Sensitivity Analysis for Longitudinal Studies with Withdrawal. PSI Conference, May 2008.

[7] Roger, J. Discussion of Incomplete and Enriched Data Analysis and Sensitivity Analysis presented by Geert Molenberghs. Drug Information Association (DIA) Meeting, Special Interest Area Communities (SIAC) - Statistics, January 2010.

[8] Carpenter J and Kenward M, 2008, Missing data in randomised controlled trials – a practical guide, Birmingham: National Health Service Co-ordinating Center for Research Methodology, www.missingdata.org.uk, accessed 07May2010.

[9] van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M. and Rubin, D. B. Fully conditional specification in multivariate imputation. Journal of Statistical Computation and Simulation, 2006, 76, 1049–1064.

[10] Lavori, P.W., Dawson, R., and Shera, D. A Multiple Imputation Strategy for Clinical Trials with Truncation of Patient Data, Statistics in Medicine, 1995, 14, 1913–1925.

ACKNOWLEDGMENTS

We would like to thank Quintiles Inc. for encouraging and supporting this work as well as conference participation. We would like to sincerely thank James Roger, Gary Koch and Kathleen Lamborn for numerous helpful discussions.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Bohdana Ratitch
Quintiles Inc.
100 Alexis-Nihon Blvd., Suite 800
Saint-Laurent, Québec, Canada, H4M 2P4
E-mail: Bohdana.Ratitch@quintiles.com