ABSTRACT

A common problem in longitudinal clinical trials is missing data caused by the withdrawal of some patients from the study before completion (that is, drop-outs). Drop-outs, in general, yield unbalanced data with unequal numbers of measurements for each patients. Some drop-outs appear completely at random. In other cases the drop-out process depends on the response variable of interest. For the latter situation, ignoring missing data may lead to biased comparison of the treatment effects. This paper compares four different techniques of estimating the mean and standard deviation of the response variable at the end of study for some longitudinal clinical trials with large percentage of dropouts. The techniques being considered include the last observation carried forward method, the mixed effects model repeated measures method (presented in PROC MIXED of the SAS software) and the incremental methods of imputation [6] programmed in the SAS software framework. The estimation methods were compared on simulated data resembling diabetes clinical trial data.

INTRODUCTION

Comparison of drug treatments involves longitudinal measurements of efficacy variables. When individual elements of longitudinal data are missing, the data are “incomplete”. The incompleteness can have different patterns and different causes. Missing values occur due to a drop-out process if the missing-data process induces a monotone pattern of missing values, i.e. a patient has observations up to a time point and none after that; otherwise the missing values are intermittent [1]. In many cases the drop-out processes possess the property that the missing data are missing at random (MAR) [2,3,4] (i.e. the drop-out process depends on the observed measurements ). A special case of MAR drop-outs is the MCAR class when the missing data are missing completely at random (i.e. the drop-out and measurement processes are independent ). A very important drop-out process falling into the MAR class is the process of withdrawal from a trial due to lack of efficacy.

A statistical analysis can be biased if incomplete cases are excluded from the analysis. The intent-to-treat (ITT) principle requires that all cases, complete or incomplete, should be included in the analysis.

There are two main approaches dealing with incomplete cases: the imputation-based approach (the missing values are filled in and the resultant completed data are analyzed by standard methods) and the mixed effects model repeated measures approach (presented in PROC MIXED of the SAS software) [5]. The final goal of both approaches is estimating parameters of a sample distribution (e.g., mean, variance) at different time points of the study.

This paper is concerned with comparison of several estimation techniques applied to incomplete longitudinal data sets with MAR drop-outs. The data sets are simulated to resemble time behavior of HbA1c in diabetes clinical trial. The distance between adjacent time points of measurements varies from 4 to 6 weeks. The missingness mechanisms employed resemble the process of withdrawal from trials due to lack of efficacy. The start of missing values for each patient depends on the most recent preceding value of HbA1c measurement. If this value exceeds a specified threshold (in general, time dependent) the patients is withdrawn from the study; otherwise, the patients remains in the study until the next measurement. The percentage of missing values in a data set depends on values of the threshold at time of measurements.

The techniques used for estimating means and standard deviations include the last observation carried forward (LOCF) method, two incremental methods of imputation (the incremental mean and incremental regression methods [6]) and the mixed effects model repeated measures method (presented in PROC MIXED of the SAS software) [5].

METHODS

LAST OBSERVATION CARRIED FORWARD

Whenever a value is missing, the last observed value is substituted. The technique is typically applied to drop-out patterns of incompleteness.

INCREMENTAL METHODS

We assume that an outcome variable $Y$ is measured for subject $i = 1,...,n$ in the study at occasions $j = 1,...,k$. The measurements are denoted by $Y_{ij}$. Let $D_{ij}$ be the increment of $Y$ for subject $i$ from time point $j$ to time point $j+1$. Next, let $Y_{ij}$ and $Y_{i}(j)$ be observed and missing values of $Y$, respectively. Similarly, let $D_{ij}$ and $D_{i}(j)$ be observed and missing increments, respectively. Denote by $D_{ij}$ the sample mean of observed increments $D_{ij}$ at time point $j$. Suppose that values $Y_{ij}$ are all observed.

The incremental mean method creates two matrices: a basic imputed matrix and an uncertainty matrix. To build the basic imputed matrix $B$ the unknown increments $D_{ij}$ are prescribed to be equal to $D_{ij}$ and, next, the missing values $Y_{ij}$, are replaced by

$$Y'_{ij} = Y_{ij} + D_{ij}$$

where values $Y_{ij}$ are observed or calculated on the previous step. The elements $B_{ij}$ that coincide either with $Y_{ij}$ or with $Y'_{ij}$ (calculated step by step as specified above) create the basic imputed matrix $B$. The approximate mean values of $Y$ at each time point $j$ will be calculated as the sample mean $B_{ij}$ using the $j$th column of matrix $B$. The columns of matrix $B$ can be used to calculate the variances of $Y$ at each time point. However, these calculated values $V_{ij}$ underestimate the real variances due to the lack of uncertainty in (1) and will be named the partial variances.

To induce an additional uncertainty it is proposed to build an uncertainty matrix $V$. Let $U_{ij}$ be the sample variance of the set $\{D_{ij}\}$ for fixed $j$. Let $R_{ij}$ be equal to 1 if $Y_{ij}$ is observed and 0 otherwise. Then the elements $V_{ij}$ of matrix $V$ are calculated as follows:

$$V_{ij} = 0; \hspace{1cm} V_{ij} = (V_{ij} + U_{ij}) (1 - R_{ij}) \hspace{1cm} \text{for} \hspace{0.5cm} j = 2,\ldots,k .$$

Therefore, if $Y_{ij}$ is observed then $V_{ij} = 0$ (there is no uncertainty). Otherwise, an additional variance (as an uncertainty measure) is accumulated over time until time point $j$ for each subject $i$. The pooled additional variance $V_{ij}$ at time point $j$ is defined as the average of $V_{ij}$ over all the subjects. The total variance $V_{ij}$ at time point $j$ is defined as

$$V_{ij} = V'_{ij} + V_{ij}.$$
\[ D_{ij} = D_j + A_j (Y_{ij} - Y_j) + \varepsilon_{ij} \]  

(4)

where \( D_j \) and \( Y_j \) are the means of samples \( \{D\}_j \) and \( \{Y_i\}_j \), respectively, over \( j \) and \( \varepsilon_i \) are errors.

For the observed data a counterpart of model (4) takes the following form

\[ D_{ij}^* = D_j^* + A_j (Y_{ij}^* - Y_j) + \varepsilon_{ij}^* \]  

(5)

where \( Y_{ij}^* \) is defined as the mean of sample \( \{ Y_i \mid R_{ij}=1 \} \) for fixed \( j \) and \( \varepsilon_i^* \) are errors.

Coefficient \( A_j \) is estimated from (5) using the least-squares method. Next, the missing values \( Y_{ij^*} \) are replaced by values \( Y_{ij^*}^* \), introduced by the following recurrent scheme:

\[ Y_{ij^*}^* = Y_{ij}^* + D_{ij^*}^* \]

(6)

After obtaining the basic imputed matrix \( B \) comprised of both \( Y_{ij}^* \) and \( Y_{ij^*}^* \) and evaluating the time point means, the calculation of the partial variances \( V_{ij} \) is completely similar to that in the incremental mean method. Consider the calculation of the corresponding uncertainty matrix \( V \). The elements of \( V \) are evaluated using the same recurrent formula (2) as in the previous method. However, evaluation of \( U_j \) is now based on the errors of model (5):

\[ \varepsilon_{ij}^* = D_{ij}^* - [D_j^* + A_j (Y_{ij}^* - Y_j^*)]. \]  

(7)

Now \( U_j \) is defined as the sample variance of set \( \{ \varepsilon_{ij}^* \} \) for fixed \( j \).

The last step of the method – calculation of the total variance is carried out using the same formula (3) as in the incremental mean method.

Note that two different means of \( Y_{ij} \) over \( i \) are used in this algorithm: \( Y_{ij}^* \) for calculation of \( A_j \) and \( U_j \), and \( Y_j \) in the recurrent scheme (6).

SIMULATION

DATA SET SIMULATION

The first step in the simulation is to create longitudinal data sets that resemble time behavior of HbA1c in a diabetes clinical trial. As an example we consider such a trial with six time points of measurements starting from baseline. The first three time increments are equal to four weeks, the next two are six weeks. We assume that the baseline distribution of HbA1c can be approximated by the normal distribution with mean of \( m_0 = 9 \) and variance of 1. The mean of HbA1c used in simulations is specified to increase by \( d_1, k=1,...,5 \), for 4th time increment, where \( d_1=0.3, d_2=-0.3, d_3=-0.5, d_4=-0.3, d_5=-0.2 \) (8)

This scenario is very approximate but, nevertheless, reflects a tendency for some diabetes drugs: a small increase in HbA1c mean value at the start, the steepest decrease of mean after approximately two months of treatment and a smaller mean decrease after that. The total decrease of HbA1c mean from baseline is specified to be one unit. The scenario of HbA1c changes is slightly different from the scenario of [6] and presents the behavior for another drug.

Each increment of HbA1c for a patient is prescribed to be the sum of a normal random variable with the standard of 0.5 and a term reflecting the process of titration (i.e. dose escalation with a set of predetermined dose levels).

Note that the random increment of the primary variable (HbA1c) increases the variance of it at the next time point. However, the titration term acts in opposite direction and keeps the variance at almost the same level. We consider below two cases: with and without titration.

MISSINGNESS MECHANISM

The missingness mechanisms employed resemble the process of withdrawal from trials due to lack of efficacy. The start of missing values for each patient depends on the most recent preceding value of ha1c measurement. If this value exceeds a specified threshold (in general, time dependent) the patients is withdrawn from the study; otherwise, the patients remains in the study until the next measurement. The percentage of missing values in a data set depends on values of the threshold at time of measurements.

Mathematically, the criterion of withdrawal used in the current simulation is the following:

\[ Y_j > m_0 + d_j + T_j \]  

(9)

where \( m_0=9, d_j \) are determined by (8) and \( T_j = 0.75 \).

Condition (9) is different from the criterion used in the simulation study [6] and presents another type of the withdrawal criterion.

RESULTS OF SIMULATION

Here we consider the results of simulation for two types of the initial (complete) longitudinal data sets (with and without titration) presented in the subsection “Data set simulation”. The number of observations is chosen to be in three versions: \( n = 30, n = 60 \) and \( n = 120 \). The missingness mechanism described above is implemented to create approximately 50% of missing data at the sixth time point. Four techniques presented in section “Methods of imputation” are compared on the base of preciseness of the mean and standard deviation estimations at the sixth time point. Statement Repeated in Proc MIXED used option type=un. For all other types the results were much worse.

To make the comparisons, we consider the bias and mean square error (MSE) of the mean and standard deviation for each estimation method involved in comparison.

The following four tables present the simulation results. Besides, the similar errors of the means and standard deviations of observed data sets are also presented. First three tables are concerned with the original data sets without titration which means that all the increments are completely random. However, the missingness mechanisms are not MCAR. The tables 4 to 6 give the results for the original data sets with titration. The missingness mechanism is indicated in each table.

| Table 1 |
|---------------------------------|-----------------|
| **Properties of data sets:**    | **Mean**        |
| - Without titration             |                 |
| - N=30                          |                 |
| - Missing: 50%                  |                 |
| **Method**                      | **Mean**        |
| Observed                        | -1.03           |
| LCOF                            | 0.46            |
| Proc MIXED: Type=un             | -0.05           |
| Incremental mean                | -0.03           |
| Incremental regression          | -0.02           |

| Table 2 |
|---------------------------------|-----------------|
| **Properties of data sets:**    | **Mean**        |
| - Without titration             |                 |
| - N=60                          |                 |
| **Method**                      | **Mean**        |
| Observed                        |                 |
| LCOF                            |                 |
| Proc MIXED: Type=un             |                 |
| Incremental mean                |                 |
| Incremental regression          |                 |
The results presented in the tables show that, for some longitudinal data sets modeling diabetes clinical trial data with different numbers of patients (30 to 120) and a large percentage of missing values due to non-responders, the incremental methods give, in average, more precise estimations (as measured by MSE) of the means and standard deviations than the other estimation methods considered in the paper. However, the bias of the mean estimation in the mixed effects model method is smaller for the case of titration than the incremental mean method. The LOCF method (widely used in clinical trials) is less precise even than the observed data approach.

**CONCLUSION**

The paper is concerned with comparison of several imputation and estimation techniques applied to incomplete longitudinal data sets with MAR drop-outs. The data sets are simulated to resemble time behavior of HbA1c in diabetes clinical trials. The missingness mechanisms employed resemble the process of withdrawal from trials due to lack of efficacy. The imputation techniques being compared include the last observation carried forward (LOCF) method, the mixed effects model repeated measures method and two incremental methods (the incremental mean and incremental regression methods). The results of simulation presented in the paper show that for different numbers of patients (30 to 120) and a large percentage of missing values the incremental methods give, in average, more precise estimations of the means and standard deviations (as measured by MSE) than the other estimation methods considered in the paper.
REFERENCES

CONTACT INFORMATION
Your comments and questions are valued and encouraged.
Contact the author at:
Naum Khutoryansky
Novo Nordisk Pharmaceuticals, Inc.
100 College Road West
Princeton NJ 08540
Work Phone: 609-987-5812
Email: nakh@nnpi.com
Web: www.novonordisk-us.com