PhUSE 2015

Paper SP05

Facing Your Fear of the Unknown – Handling Missing Data

Amanda Reilly, ICON plc, Dublin, Ireland

ABSTRACT

Ideally, data collected during a clinical trial would be complete, with subjects not missing any visits or dropping out. However, this is never the case in reality. In fact, missing data is a common and serious issue for clinical trials as it can limit the conclusions drawn or camouflage safety issues with the treatment. The focus of this paper is to describe various techniques used to handle missing data, to help statisticians and statistical programmers make an informed choice on the most suitable technique to use for the data being analyzed.

The paper is targeted at statisticians and statistical programmers within the pharmaceutical industry who have a basic understanding of statistical analysis.

INTRODUCTION

Missing data is a common and serious issue for clinical trials, so it is important for statisticians and statistical programmers to be aware how to handle missing data, as well as the consequences of ignoring missing data. This paper identifies what is missing data, possible reasons for missing data, as well as when and how they should be handled.

Most statisticians and programmers are familiar with the Last Observation Carried Forward imputation technique, but is it really the Holy Grail of imputation methods? Although not an extensive review of all available methods, this paper describes this and various other techniques that can be used to handle missing data.

WHAT ARE MISSING DATA?

Missing data are defined by Little et. al[1] as "values that are not available and that would be meaningful for analysis if they were observed". Ideally, data collected during a study would be complete, with no observations missing, but this is rarely the case in reality. In fact, the issue of missing data is a very common and serious one for clinical trials and it can be a complex issue to address. There are knock-on effects such as limited conclusions or camouflaged results. For a successful and accurate analysis, it is paramount that missing data are handled correctly, as the approach taken may affect the power and results of the trial.

In a clinical trial, early discontinuations from the trial are the main source of missing data. There can also be many other reasons for missing data, including a data entry error (measurements taken but not recorded properly), relocation of a subject, the death of a subject (for reasons unrelated to the underlying condition or treatment), drug intolerance, drug effectiveness or ineffectiveness, or a subject being lost to follow up.

Do all missing data matter? The impact of the missing data may depend on the reason for it being missed. For example, if a clinical trial subject had considerably worsened, they may have been unable to attend a scheduled visit, or may have withdrawn from the trial completely. Had they managed to attend the visit, or continue in the trial, the measurements from the missed visit may have impacted their overall results due to particularly high or low measurements recorded. Consider, also, a subject who decided not to answer a particular question on their questionnaire. Could this missing value be as a result of the subject's embarrassment, or something they were afraid to admit? If so, perhaps the recording may have been of considerable interest in the analysis, had it been observed.

WHY SHOULD MISSING DATA BE CONSIDERED?

The power of a trial is its ability to reliably detect and measure the difference between the treatment and control groups, for an effective treatment. The aim is to avoid an incorrect conclusion that the treatment is ineffective (known as a type II error). The statistical power of a trial increases along with its sample size, or as the variability of the outcome reduces, and therefore it is important to use as many of the randomized study participants as possible. Excluding subjects with missing values from the analysis will result in a reduced sample size, and therefore a reduced statistical power for the trial, making it more difficult to detect a treatment effect, if it is truly there.

Bias in the estimation of the treatment effect is the inclination to be in favour of the treatment group over the control group, usually caused by ignoring the uncertainty around missing values when estimating the standard errors. Bias is a very important concern around missing data, and the risk of this bias depends on the relationship that the missing data have with the treatment and outcome.

MISSING DATA REGULATORY GUIDELINES

For clinical trials in the pharmaceutical industry, the European Medicines Agency (EMA) was established in 1995 to protect and promote public and animal health, through the evaluation and supervision of medicines for human and veterinary use. The EMA has seven scientific committees carrying out scientific assessments, one of which is the Committee for Medicinal Products for Human Use (CHMP). This committee are responsible for preparing opinions on questions concerning medicines for human use. In 1998, the CHMP (known as Committee for Proprietary Medicinal Products, CPMP, at the time) released a document known as "ICH Topic E 9 - Statistical Principles for Clinical Trials"[4]. It defines missingness as "both the existence of missing data and the mechanism that explains the reason for the data being missing". It confirms that missing data are a potential source of bias in a clinical trial, and therefore stresses the importance of minimising missing data in a trial. It highlights the role of the protocol to specify procedures that proactively plan to minimize the extent of missing data in the trial design.

ICH E9 also indicates that missing data may be compensated for, using imputation techniques, as long as strategies are described and justified in the protocol, along with any underlying assumptions of the data being made. It also indicates that a trial is regarded as valid if the methods of dealing with missing values are sensible and pre-defined in the protocol, and the methods may be refined in the statistical analysis plan of the trial.

As ICH E9 only partially covers the topic of missing data, the CHMP (still known as CPMP at the time) released a document named "Points to Consider on Missing Data"[5] in 2001, which was replaced in 2010 with a "Guideline on Missing Data in Confirmatory Clinical Trials"[6]. Again, this document stresses that "it should be the aim of those conducting clinical trials to achieve complete capture of all data from all patients, including those who discontinue from treatment". It states that interpreting the results of a trial with a substantial proportion of missing values is problematic as "the uncertainty of the likely treatment effect can become such that it is not possible to conclude that evidence of efficacy has been established". The guideline says that "an appropriate analysis would provide a point estimate that is unlikely to be biased in favour of experimental treatment to an important degree (under reasonable assumptions) and a confidence interval that does not underestimate the variability of the point estimate to an important extent". The number, timing, pattern and reason for missing values in previous trials should also be investigated, to help identify additional actions that may reduce the amount of missing data in further trials.

The guidelines also note that "just ignoring missing data is not an acceptable option when planning, conducting or interpreting the analysis" of a randomized clinical trial aiming to assess if a treatment effect observed in a previous randomized trial is real or important, as this type of trial "should estimate the effect of the experimental intervention in the population of patients with greatest external validity".

WHEN SHOULD MISSING DATA BE ACKNOWLEDGED?

We cannot know the reason for each missing value, so do all missing data need to be accounted for and acknowledged? How do we decide what missing data are important to consider? The key factor is the relationship between missingness, treatment assignment and outcome. The CHMP guidelines show how the relationship is classified by 3 different terms; MCAR, MAR, and MNAR.

As per the CHMP guidelines, the missing data are referred to as Missing Completely at Random (MCAR) when "the probability of an observation being missing does not depend on observed or unobserved measurements" for the dependent variable (conditional on the covariates in the model). In simpler terms, the presence or absence of a value is unrelated to value that may have been observed. The guidelines give an example of data MCAR as subjects who move to another city for non-health reasons; they are considered a random and representative sample from the total study population.

Missing data are referred to as Missing at Random (MAR) when "conditional on the covariates in the model, the probability of an observation being missing depends only on observed measurements". This implies that post-dropout behaviour can be predicted from pre-dropout observed behaviours, so the response can be estimated without bias. An example of data MAR are when subjects drop out of a trial due to lack of efficacy reflected by an observed series of poor efficacy outcomes, and it would then be appropriate to impute or model poor efficacy outcomes subsequently for this subject. As explained by Burzykowski et al.[7], MAR is more easily explained by considering a patient drop-out in a longitudinal study; "Suppose that two patients share the same treatment and covariates, and exactly the same response measurements up to the point at which one drops out and the other remains. Then the missing data from the subject who drops out are MAR if they have the same statistical behaviour as the observations from the subject who remains. Under MAR, a valid analysis can be constructed that does not require knowledge of the specific form of the missing value mechanism".

Observations that are neither MCAR nor MAR, are then classified as Missing Not At Random (MNAR). CHMP define this as when "the probability of an observation being missing depends on unobserved measurements", where "the value of the unobserved responses depends on information not available for the analysis". In simpler terms, the presence or absence of the data is related to the value that may have been observed, and therefore, unobserved observations cannot be predicted without bias in the model. An example would be subjects dropping out of a clinical trial due to declining health. Although it's best not to ignore any missing data due to the effect on the power of the trial, it's particularly important not to ignore missing data of this type, because otherwise estimates would be biased since missing data are related to the outcome. If the presence or absence of the data is also related to treatment group, then ignoring the missing data would give invalid statistical tests.

As mentioned above, the relationship between missing values and the unobserved outcome variable cannot be defined with certainty, so it is impossible to know if data is MCAR, MAR or MNAR. It cannot be confirmed if assumptions are appropriate, and it is very difficult to judge if missing data can be adequately predicted from the observed data. The CHMP guidelines say that "the justifications for the methods chosen should not depend primarily on the properties of the methods under the MAR or MCAR assumptions, but on whether it is considered to provide an estimate that is acceptable for regulatory decision making in the circumstances of the trial under consideration". Due to this uncertainty, it is often best to adopt a conservative approach and assume any missing values are a potential source of bias.

In summary, bias is a very important concern resulting from missing data, and the risk of this bias depends on the relationship that the missing data has with the treatment and outcome. In particular, missing data that are MNAR (i.e. the probability of an observation being present or absent depends on unobserved measurements) should not be ignored as this will lead to biased estimates and, at worst, invalid statistical tests.

HOW SHOULD MISSING DATA BE HANDLED?

The next question, of course, is "How should missing data be handled?". There is only one completely correct way to handle missing data, and that is to prospectively ensure that there are no missing data! In reality, of course, some missing data are usually unavoidable, but it is very important to minimize missing data as much as possible in a trial. The protocol should specify procedures that proactively plan to minimize the extent of missing data in the trial design.

Incomplete data will still need to be analyzed, so the best approach to handling the missing data correctly must be chosen. Deciding how missing data should be handled can become quite complex, as options can depend on the objective of the study and the type of trial in question. As indicated in the CHMP Guideline on Missing Data[6], "there is no universally applicable method that adjusts the analysis to take into account that some values are missing, and different approaches may lead to different conclusions". How missing data are handled can have an important influence on the overall results and conclusions of the clinical trial and, since analysis methods rely on assumptions that cannot be verified, how missing data are handled might be a source of bias in itself! Efforts should be made to make reasonable assumptions about the data being analysed, in order to choose an acceptable method to handle the missing data, and conservative methods that are unlikely to favour the experimental treatment should be considered.

Subjects with missing values might be more likely to have had extreme values, depending on their reasons for not responding to a particular question, or for dropping out of the study. Therefore, the loss of these missing values may lead to an underestimate of variability in the data, and an artificially narrow confidence interval for the treatment effect. The methods used for handling missing data should, therefore, attempt to account for this uncertainty of the treatment effect in order for the confidence interval to be considered valid. If the amount of missing data is substantial, a sensitivity analysis is recommended (and, in fact, critical for a clinical trial from a regulatory perspective), to check the robustness of the analysis and examine the sensitivity of the results to the assumptions made regarding the missing data.

DELETION METHODS

The most basic and common techniques of handling missing data are deletion methods, where observations with missing data are excluded from the analysis (as discussed by Little (1992)). An example is listwise deletion (also referred to as 'complete case analysis'), where an entire record is removed from the analysis if it contains any missing value.

These methods are often favoured for their simplicity and ease of use. Of course, the study will have a reduced statistical power if these methods are used, due to the reduction in the size of the sample used. If the data are missing at random (i.e. the probability of an observation being missing depends only on observed measurements) or missing not at random (i.e. the presence or absence of the data is related to the value that may have been observed) then deletion methods may also result in biased estimates.

The CHMP Guideline on Missing data[6] advises that this technique is unsuitable as the primary analysis in a confirmatory trial (one which aims to assess if a treatment effect observed in a previous randomized trial is real or important), as it violates the intention to treat principle and is subject to bias. This method should only be considered in exploratory studies, or as a sensitivity analysis in a confirmatory trial (to demonstrate robustness of conclusions).

IMPUTATION METHODS

In some cases, it's appropriate to represent the missing values with 'imputed' values. Imputation is the process of missing values being replaced with a value, making use of the available data to best estimate the missing values. Once all missing values have been replaced, the observed and imputed data are set together and the 'complete' dataset is then analysed as normal. The theory of imputation is constantly developing and growing as more research attempts to find better ways to handle missing data without introducing a large amount of bias.

IMPUTATION TECHNIQUES

There are various techniques for imputing data. Choosing a suitable technique depends on various aspects of the data being imputed that the analyst should consider, such as the form of the missing data (numerical, binary or categorical), the objective of the study and the structure and design of the study (i.e. is there a nesting structure to consider?).

SINGLE IMPUTATION

Single imputation techniques involve replacing each missing value with a single value. The single imputation method of Last Observation Carried Forward (LOCF) has been widely used in clinical trials. The method is suitable for longitudinal studies with multiple post-baseline visits or measurements and involves imputing a missing post-baseline value by carrying forward the most recent observed post-baseline result, if one exists. If a subject drops out of a study early, the most recently observed post-baseline result is used to impute the remaining measurements that were expected. Advantages of this technique are that it is simple to understand and apply, but there are various aspects (such as windowing) to consider in the implementation of this technique, as detailed by Jain and Pandya[2].

So, is it really the Holy Grail of imputation methods? Little et. al[1] claim that this technique is overused and leads to biased treatment effects. Ting and Brailey[3] also claims that LOCF gives a biased point estimate and biased variance, and the CHMP Guidelines on Missing Data[6] confirm that LOCF only produces an unbiased estimate of the treatment effect "under certain restrictive assumptions". Since the underlying assumption of LOCF is that subjects gradually improve over the study period, a conservative approach for missing data is thought to be assuming no change in measurements (i.e. they continue as per the most recent non-missing, post-baseline measurement). Therefore, early drop-outs of a trial are assumed to plateau after withdrawal (so post-dropout behaviour can be predicted from pre-dropout observed behaviours). However, this is unrealistic, as drop-outs are more likely to get sicker once the treatment effect begins to wear off, so this approach isn't considering the reason why the value is missing (i.e. why the subject dropped out). In fact, LOCF has been largely discredited and discouraged in favour of modelling approaches (discussed below).

Imputing with the best and worst case is a common single imputation technique for binary outcome data. Initially, the worst case scenario is assumed for the test drug, where control group subjects have a good outcome but test drug subjects do not. It is then assumed that the test drug subjects have a good outcome, but control group subjects do not (best case). Assuming the worst case for the test drug is a very conservative approach, and provides the lower limit for the treatment effect, while assuming the best case provides the upper limit for the treatment effect.

Other examples of values that can be used to impute with are the mean of the non-missing values of that variable (referred to as 'un-conditional mean' imputation, where there is no input from other variables), or the predicted values from regression analysis (referred to as 'conditional mean' imputation). Although it is a simple solution, a disadvantage of this technique is that the distribution for the variable can be distorted. Using a mean value to replace missing values leads to biased parameter estimates and underestimated standard deviations, giving confidence intervals for the treatment effect that may be too narrow. This, in turn, can lead to an increased risk of finding something significant when in fact it is not (referred to as a type I error).

In general, single imputation methods are often favored due to their simplicity. However, the drawback is that the uncertainty of the unknown missing value is not accurately reflected using these methods.

MULTIPLE IMPUTATION

The multiple imputation method attempts to resolve the single imputation issue of the uncertainty of the unknown value not being reflected, by adding variability to the analysis to reflect this uncertainty. Instead of imputing each missing value with a single value, the multiple imputation technique results in each missing value having a random sample of plausible values, and the uncertainty around the missing value is reflected in the distribution of this sample. The observed data are used to estimate the distribution parameters of this generated sample. As discussed by Gelman et al. (1998)[9], this method assumes that the data are MAR (i.e. that the probability of missingness depends only on observed data included in the model).

The multiple imputation procedure generates M copies of the original dataset, each one imputing the missing values with various values deemed plausible by the observed data. Each of these M complete datasets are then analysed separately using standard statistical methods. Finally, the M results from these analyses are combined to produce one set of inferential results with valid statistical inferences (such as confidence intervals with the correct probability coverage) that properly reflect the uncertainty in the data due to the missing values.

Using PROC MI in SAS®, a complete dataset of the M combined datasets can be created which contains the original observed values as well as imputed values in place of previously missing values. This new dataset is M times the size of the original dataset, as each row is repeated M times for the multiple imputations. This complete dataset also contains an index variable, "_Imputation_", which labels the M rows for each observation in the original dataset with values of 1 to M. For observations where no imputation is needed, the observation is repeated M times in the new dataset with each row labelled 1 to M using the imputation index variable. When imputation is needed, the

original observation is repeated M times but, in each row, previously missing values are replaced with various plausible values for that observation. The new complete dataset can then be analysed using the standard statistical procedure by adding "_Imputation_" as a "BY" variable. For example, the following code can be used to perform a t-test analysis on the dataset generated from PROC MI:

```
proc ttest data = dataset;
by _Imputation_;
class XXXX;
var YYYY; ;
run;
```

This gives M sets of results and parameter estimates, which are then pooled together to give one overall result using PROC MIANALYZE in SAS. This procedure uses the same process of combining the results, regardless of what analyses were used. The procedure reads parameter estimates (and their associated standard errors or covariance matrix) computed in the M separate analyses, and then derives a valid univariate inference for these parameters. For parameter estimates, the estimated parameter is the average parameter value across all M imputed data sets.

The syntax for PROC MI is as follows:

```
PROC MI <options>;
BY variables;
EM <options>;
FREQ variable;
MCMC <options>;
MONOTNOE <options>;
TRANSFORM <options>;
VAR variables;
RUN;
```

There are various options available to add to the procedure to impose restrictions on the imputed values generated, such as rounding the imputed values, choosing a seed for the random generation, or applying a minimum and maximum imputed value. The PROC MI statement is the only required statement for this procedure, and it is in this statement that the input and output datasets are specified.

The 'VAR' statement lists the numeric variables to be imputed. If this statement is not included, all numeric variables not listed in other statements will then be imputed. If restrictions are applied to the imputed data (such as a minimum, maximum, rounding), then a restrictive value is required for each corresponding variable listed in the 'VAR' statement. Otherwise, the restriction will be applied to all variables being imputed.

The 'BY' statement specifies the grouping used to perform the multiple imputation, and is very useful if the data is structured into clusters; the mean and variation of observed data within each cluster can be used to impute missing values within that specific cluster. Similarly, data can be grouped by treatment group, or any other variables suitable for the particular data being imputed.

As an example, the code below imputes variables XXX and YYY, based on the observed data within each treatment group specified in the variable TRT. Both variables, XXX and YYY, will be imputed with integer values, since round=1 is applied to both (since only one rounding value is specified). Variable XXX will be imputed with values between 2 and 6, while variable YYY will be imputed with values in the range 3 to 10.

```
PROC MI    data = input_dataset1
           Round = 1
           Minimum = 2 3
           Maximum = 6 10
           Out = outputdataset1;
VAR XXX YYY;
BY TRT;
Run;
```

The syntax for PROC MIANALYZE (where P_E is the point estimate and S_E is the standard error of the point estimate) is as follows:

```
PROC MIANALYZE data = input_dataset2;
    ods output ParameterEstimates = output_dataset2;
    modeleffects P_E;
    stderr S_E
RUN;
```

MODELLING TECHNIQUES

Missing data can also be handled through modelling approaches. Linear mixed models can be used to handle missing values for continuous data with a series of repeated measures of outcomes over time (MMRM; mixed-effect models for repeated measures). This is a commonly used technique for clinical trials and, under the assumption that the missing data are MAR (i.e. that subjects who drop out of a study early follow the same trajectory as those who complete the study), provides an unbiased estimate of the treatment effect. For categorical and count data, the marginal and random-effects approaches are used by generalized estimating equations and generalized linear mixed models, respectively. The option settings for these models must be carefully considered, pre-defined and justified to avoid bias, since different settings may lead to different conclusions.

However, these methods are criticized by regulators because they assume that data are MAR. However, patients who withdraw from a trial early are clearly different to those that complete the trial, as any treatment effect can be expected to last for only a short time. Therefore, the patient who dropped out from a trial treating an illness will most likely be sicker than the one that didn't, indicating that missing data are not imputed accurately and the size of the treatment effect will most likely be overestimated and biased. Even so, they are powerful methods and shouldn't be dismissed. They are acceptable under the assumption, but since the underlying assumption isn't a favourable one, we need to also use other methods that assume MNAR and check for consistency of results. If methods with different assumptions result in the same conclusion, it is good evidence that the conclusion is correct.

Pattern mixture models (PMM), as discussed by Ratitch and OKelly (2011)[10], is another way to model missing data, and can be used as a sensitivity analysis to test how much assumptions affect the trial results. This is done by imputing the missing data under prespecified assumptions and analysing the data, and then checking how results change under varying assumptions in order to test the robustness of the results. The PMM method is only applicable to longitudinal data and involves choosing cohorts based on the patterns of missing data (so subjects in a cohort share the same pattern of missing data). A link between observed and unobserved values is then specified and used to perform multiple imputation.

SUMMARY

In summary, missing data are defined as values that are not available, but would be meaningful for analysis if they were. Missing data are a common and important issue in clinical trials, and should not be ignored if the absence of the data is related to the value that may have been observed if it were not missing.

In general, there is no 'good' approach to handling missing data, other than preventing it. There are, however, various approaches that can be used, for which the underlying assumptions vary. Care is needed to choose a method that is most suitable for the data and study in question, and for which the underlying assumptions are appropriate. This choice may affect the results of the analysis and may be a cause of bias in itself. No single method works completely, and the type of missing data in question can never really be proven, as the values are unknown. Methods with different underlying assumptions about the data should be used (for example, data are MAR or MNAR), and consistency across the results should be checked. The power of the trial, the risk of bias in the estimation of the treatment effect and the CHMP guidelines should also be considered.

Deletion methods result in a reduced power for the trial and possibly biased estimates, and should only be used in a pilot study. There are various imputation techniques available for use when imputing missing data, but care should be taken when choosing a technique that is suitable for the data and analysis of the trial in question. Also note that a subject can discontinue from the study, or they can discontinue from the treatment but remain in the study. How both scenarios would be imputed should be considered. Multiple imputation methods are generally more robust than single imputation methods due to added variability which reflects the uncertainty around the missing data.

REFERENCES

1. Little et al. "Prevention and Treatment of Missing Data in Clinical Trials". Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3771340/

2. Jain and Pandya "LOCF: It's More Than Just Carrying Forward". Available at: http://www.lexjansen.com/nesug/nesug09/ph/PH03.pdf

3. Ting, N. and Brailey, A. (2012). "Why do people use locf? or why not?". Available at: http://www.doc88.com/p-517461081967.html

4. EMA and CPMP (1998). "ICH Topic E 9 - Statistical principles for clinical trials". Available at: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002928.pdf

5. EMA and CPMP (2001). “Points to Consider on Missing Data”. Available at: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003641.pdf

6. EMA and CPMP (2010). “Guideline on Missing Data in Confirmatory Clinical Trials”. Available at http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2010/09/WC500096793.pdf

7. Burzykowski, T., Carpenter, J., Coens, C., Evans, D., France, L., Kenward, M., Lane, P., Matcham, J., Morgan, D., Phillips, A., Roger, J., Sullivan, B., White, I. and Yu, L.M. (2010). “Missing data: Discussion points from the PSI missing data expert group”. Available at: http://onlinelibrary.wiley.com/doi/10.1002/pst.391/epdf

8. Little, R.J. “Regression with missing x’s: A review” (1992). Available at: http://www.jstor.org/stable/2290664?seq=1#page_scan_tab_contents

9. Gelman et al. (1998). “Not asked and not answered: Multiple imputation for multiple surveys”. Available at: http://gking.harvard.edu/files/not.pdf

10. Ratitch, B. and OKelly, M (2011). “Implementation of pattern-mixture models using standard sas/stat procedures”. Available at: http://pharmasug.org/proceedings/2011/SP/PharmaSUG-2011-SP04.pdf

RECOMMENDED READING

1. The MI Procedure. Available at: http://www.ats.ucla.edu/stat/sas/v8/miv802.pdf

2. Gullion et al. “A Method of Using Multiple Imputation in Clinical Data Analysis”. Available at: http://www.lexjansen.com/pnwsug/2008/ChristinaGullion-MultipleImputation.pdf

CONTACT INFORMATION

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

Author Name: Amanda Reilly
Company: ICON Clinical Research Ltd
Address: South County Business Park, Leopardstown
City / Postcode: Dublin 18
Email: Amanda.Reilly@iconplc.com
Web: http://www.iconplc.com/