CLINICAL INVESTIGATION

INTERVIEW SERIES

Clin. Invest. (2012) 2(1), 19–24

Missing data in clinical trials: a data interpretation problem with statistical solutions?

"Novel methodology is not the main issue. Understanding the limitations that missing data place on interpretation of data is primary, followed by the need to specify the question at hand."

James Roger speaks to Laura Harvey, Assistant Commissioning Editor.
James Roger recently retired from the Research Statistics Unit at GlaxoSmithKline. He is honorary professor at the London School of Hygiene and Tropical medicine and an honorary member of Statisticians in the Pharmaceutical Industry. His research interests include mixed models missing data and the use of likelihood-based methods of inference.

Q How would you assess the importance of a uniform approach to dealing with missing data in terms of 'correctly' interpreting trial data and the impact of misinterpreted data?

For the public and specifically the potential recipient of any therapy, it is important that any individual trial and the collection of data from several studies are interpreted in a way that predicts future usage. It's the role of regulators to represent that interest and is an obligation on the industry to supply information that allows such an interpretation.

Q What would you say is the best way to go about supplying this information?

This is best served by clearly stating the purposes of the trial and by understanding the potential impact of any unobservable data on predicting the outcome during future usage. For instance, individuals will withdraw for several reasons, some unrelated to the trial, some associated with poor outcome and some related to extremely good outcome.

In the process of synthesizing data, many assumptions are made about the representativeness of the trial and the summary measures taken from that trial, such as patient population, background therapy, co-medication, choice of end point and length of trial. Amongst those assumptions about representativeness is the extent to which those who withdrew from the trial are different in potential outcome and also response to treatment from those who remained throughout. Any uniform approach needs to be able to reflect the range of possible assumptions. As such there may not be a single 'correct' interpretation. It will depend upon assumptions and these assumptions will often not be testable.

Q In light of these possibly 'untestable' assumptions, what do you think is most crucial in terms of strategy to deal with interpreting the missing data?

From an industry perspective, it is important to develop an agreed approach, so that future trials can be designed, run and analyzed along these principles. Compared to current practice, additional postwithdrawal data may need to be collected and sample sizes may need to be increased to allow for additional perceived variability in outcome measures. That is, trials may need to be more precise

James Roger
Livedata (UK) Ltd, Chinnor Hill Lodge, Chinnor Hill, Oxfordshire, O39 4BQ, UK
E-mail: james@livedata.co.uk

ISSN 2041-6792
10.4155/CLI.11.147

in the traditional statistical sense so as to allow for uncertainty about the impact of missing data.

Q In the absence of one 'perfect' statistical method to apply in the event of missing data, what is your opinion on the need for a universal, principled, approach to handling the missing data perfectly?

Undoubtedly we need to move to a position where there are agreed principles for handling all types of missing data. I suspect that the 'how?' part (statistical methodology) will be the easy part. Indeed we already have some agreement on best practice from a statistical point of view once the estimand is defined and certain assumptions have been agreed about the withdrawal process.

Q What will the difficult part be?

The difficult part will be developing a consensus of what readout is required from each individual trial. Developing principles for this will be difficult. For instance, in a long-term trial in amyotrophic lateral sclerosis patients, where mortality rates may be as high as 25% per annum, how do you measure disease progression? How do you place death beside progression of symptoms and loss of mobility? It may well be the case that within a specific type of trial, for a specific end point there may evolve standard approaches (for these specific criteria) to define the outcome measures and to interprete the impact of missing data.

Q Can you give an example?

Well, for instance, in trials on Alzheimer's disease using Alzheimer's Disease Assessment Scale-cognitive subscale as outcome, a standard approach to handling missing data is likely to emerge. Perhaps it will be based on mixed-effects model repeated measures as primary and a series of specific sensitivity analyses.

Q How do you think the recent European Medicines Agency and National Academies guidelines on handling missing data will be interpreted and followed in the analysis of clinical trials?

We are already seeing the impact of the Committee for Medicinal Products for Human Use (CHMP) guidance in the regulatory responses for guidance within the European Medicines Agency region. Specific reference to it is being made and the whole flavor of responses reflect the mood of the document; 'conservative' primary analyses and an interest in effectiveness rather than efficacy. The document officially came into effect as a guideline at the beginning of 2011, but we saw it being referred to in regulatory response before that date.

On the other hand the National Academies (NA) document is a committee response to a request from the US FDA. It is a much larger and extensive document with much more detail and examples. Sometimes the messages are mixed, as one might expect from a committee. Some of the ideas, while interesting in stimulating discussion, may be slightly impractical, reflecting perhaps the bias towards the academics on the committee. The most important message from this document is the concept of an estimand and I am sure that this term will make its way into any future FDA guidance.

Q Do you see differences between the documents as potentially adding to the confusion over how to deal with missing data?

The documents are very different in form as they have different purposes and different audiences. The CHMP document is a fully fledged guidance that tries to deal with principles rather than with detail [1]. Also there is an apparent desire to not 'give ground' that might need to be 'clawed back' in some future version. It is my impression that the regulatory statisticians do not see the document as the final gold standard, so further revisions should be expected. Guidance *per se* always tends to veer on the conservative side.

On the other hand, the NA document is far more of a discourse and as such can explore less conventional areas [2]. But there is the same underlying message in both documents. First, decide what it is that you are trying to estimate (the estimand) and then second, document and understand what happens to those who withdraw.

Q The main aim when dealing with missing trial data is to understand what happens to those who withdraw, that much is clear. However, would you say that differences in the recommended practical approaches as to how to do this are where problems arise and how do you see the differences in the two documents contributing to this?

The major difference I see is that the CHMP guidance looks towards measuring effectiveness in an intention to treat (ITT)-like fashion, whereas the NA document seems to be more closely aligned to analyses that answer an on-treatment question. The CHMP document recommends the possible use of 'retrieved dropout' data. Here they are looking to answer an ITT-like question, including in the outcome measure the response of those who have come off treatment and gone onto some alternative regimen. On the other hand, the NA document, when it calls for the collection of postwithdrawal data, sees it as a way for understanding the withdrawal process better. That is, they want to guess what would have happened had those subjects stayed on treatment. Apparently they wish to continue to answer the on-treatment questions but allow for any bias introduced by withdrawal selection. This potential distinction in the possible estimand is a very interesting and important one that has only recently received suitable discussion.

Methodology

Q In terms of implementing a uniform approach to trial analysis, what would you say to the argument that, in order for uniformity across clinical trials, all analyses should rest on an agreed assumption about the unobserved data and the reason that they are missing? What are the difficulties in applying this assumption to all types of trials and analyses?

The phrase 'missing data' can give the wrong impression. In many cases the data were never there. The phrase suggests that the there is a uniformity to the subjects who withdraw, that is not valid. Subjects withdraw for a series of different reasons; some because they are doing well, others because they are doing badly. Sometimes the reason for withdrawal will be captured, sometimes the reason will be missing. One cannot expect to have a uniform agreed assumption. Data might be sensibly analyzed as missing at random conditional on a covariate in one trial. But when that covariate is not collected in a similar trial run by a different investigator but in the same scenario, it will lead to a situation where a missing – not at random – analysis is required.

Q How do you think regulators should decide whether or not a particular (analysis) method is acceptable?

It must come down to whether the method is likely to provide a fair estimate. There are two major parts to fairness:

- Does it answer the required question?
- Is it biased by the way that certain information is not available?

It should not simply be about whether the method is 'conservative'. The same method will behave conservatively on average in some settings and liberally on average in other settings. And this begs the key question because one is left with the problem of understanding which of those situations one is in. For instance, last observation carried forward will be conservative where withdrawals are due to lack of efficacy in a degenerative disease leading to more dropouts in the placebo arm, while it will be liberal in the same scenario if the surfeit of withdrawals is in the active arm resulting from unacceptable side effects. It may not be obvious before the trial is run which of these two is more likely. Whether a method is acceptable has to rely on the scenario and has to reflect the risk associated with untestable assumptions. However it remains important to specify those assumptions up front before the trial is run.

Q Should the scientific community be more open to novel methodology and what is the best way to introduce new methods to data analysis?

Novel methodology is not the main issue. Understanding the limitations that missing data place on interpretation of data is primary, followed by the need to specify the question at hand. One practical limitation is the availability of easy to use and understand software for implementing suitable methods. We have made available on the London School of Hygiene and Tropical Medicine website a SAS macro that implements a series of sensitivity analyses using multiple imputation and missing not at random models based on pattern mixture models. An open research area is to define suitable sensitivity analyses for discrete data such as counts of exacerbations in coronary obstructive pulmonary disease trials.

Q What happens if different analyses give different results in terms of estimated effects or in terms of statistical significance?

First one needs to understand whether two competing analyses are trying to estimate the same thing and answer the same question. If not, then this may simply explain the divergence. If they are estimating the same thing, then one must look to the assumptions behind the approaches. To what extent might one expect them to lead to different conclusions? The pattern mixture approach allows one to summarize the outcome measure across time both before withdrawal based on actual data and after withdrawal based on the mean of predicted data under the assumptions for the differing methods. The approach can even be used with data-based rules such as last observation carried forward. The estimated effects for each approach will be some form of average across these patterns. So a difference in the average implies there must be a

difference somewhere within one or more of the patterns. This leads one back to understanding which set of subjects are generating the distinction. This in turn allows one to identify the aspects of the methods that lead to the difference in interpretation. Stratifying the data by the pattern of withdrawal is a powerful tool.

Q What role does retrieved dropout information play in minimizing the impact of unobserved data?

Retrieved dropout allows one to answer a series of effectiveness questions, but it does not directly help in answering on-treatment types of question. Information collected after withdrawal can help understand the withdrawal event but it is secondary and does not allow full recovery.

In planning a trial with collection and analysis of retrieved dropout data, a series of new challenges appear:

- What therapy is allowed in the off-treatment treatment period?
- Will there be any restrictions?
- Does collecting retrieved dropout data limit the protocol in this way?
- Understanding retrieved dropout data requires us to collect accurate information about concomitant medication;
- Is this possible in a long term study?
- How do we use it?

Using such retrieved dropout data is a difficult statistical challenges. It certainly is not an easy solution and does not 'minimize the impact of missing data' although it may help with a *de facto* type of estimand.

I am intrigued by the Regulator's interest in retrieved dropout as I believe it reflects their interest in *de facto* estimands rather than *de jure* estimands. I believe that it also reflects their reticence towards some mixed models for repeated measures-type analyses.

Q Does this cause some concern within the industry?

Yes, some in the industry are worried by this as they feel it will make the barriers even higher to successful licensing. We have seen that for some analyses, as one moves from *de jure* to *de facto*, the estimated treatment difference becomes smaller in absolute value; however, the standard error of the difference also shrinks leading to a similar level of significance. That is, the treatment is still proven to work (better than zero) but the predicted marginal effect is likely to be smaller.

Design

Q Given the low cost (compared with trial costs as a whole) for extra work/preparation at the trial design and analysis stages, why do you think there is still difficulty in this area and what role do you think regulatory bodies can play in providing clinicians with a 'go to' list of questions to approach missing data?

First, much of the current methodology looks to the nonstatistician like a black box. On the other hand a single imputation approach is understandable, even though it may have strange statistical properties. Methods that simply modify the data have appeal. It is important that statisticians take care to describe valid statistical approaches in a friendly but intelligent way. I believe that pattern-mixture models and multiple imputation facilitate this selling of the message. Subjects are stratified by when they withdraw from the trial. Importantly this is a post-randomization stratification, so we may expect differing numbers allocated to each pattern in different arms of the trial. The treatments drive the withdrawal patterns. We observe what happens within each of the patterns.

Sensitivity analyses will be crucial in any best practice. These open up a trial for multiple interpretations. There is a feeling in the industry that the regulators will take the worst case scenario. So it is perceived as important to provide as few scenarios as possible. The more hurdles, the more chances of a fall. This attitude is not in anyone's interest. But it will take a while for the industry to have confidence that supplying a wide range of sensitivity analyses improves the chances of a positive review rather than the opposite.

Q Do you think this is an effect of the pressure on pharmaceutical companies to produce results as quickly as possible?

Yes, there is always a rush to get results out and no sponsor wants an analysis where the computations might not converge or software options, such as type of algorithm, need to be tuned. So the industry is going to be wary of complex analysis methods and multiple alternative sensitivity analyses.

Q How can clinical trials be adapted/structured at the design stage to accommodate for missing data?

At the design stage it is important to identify measurements that are both associated with outcome and are associated with withdrawal. By conditioning on these we make the missing at random assumption a better approximation. Then by matching the analysis to the question asked and also by making a missing at random assumption we can derive robust conclusions.

Phase III trials should be designed to answer a question in terms of the expected value, population margin, of some utility function, often an outcome measure. Discussing what form that utility function might take, can be useful for a development team. The handling of missing data should then reflect that utility function and the required margin in terms of population. I feel this approach could also be useful in discussion with regulators. However, most pharmaceutical statisticians feel happier to just discuss the choice between individual standard analysis methods.

Q In your opinion how do you know when a certain amount of missing data renders a trial invalid? Do you think that there should be guidelines in place such that researchers know when their collected data are not suitable for use?

I would be very worried if we had a guidance that trials with more than say 30% missing data are classed as unanalysable. The important thing is the impact of the missing data on the conclusions. With everything else held constant the impact will increase with the amount of missing data. But the importance of 10, 20 or even 30% of missing data will depend upon a series of other things; the reasons for the withdrawals, the imbalance in the frequency of those withdrawals between the arms and the timing of those withdrawals.

Q What would you suggest for inclusion in such guidelines and how do you think they should be implemented?

Guidance needs to indicate how to investigate uncertainty about assumptions. By measuring the impact of missing data we can see whether it influences the conclusions to such an extent that they are at risk. I suggest that this is done by proposing alternative postwithdrawal trajectories for different types of withdrawal and identifying the impact of these using multiple imputation. But alternative statistical methodology may be able to do something similar. The main issue is to be able to measure the potential impact of something at which we can only guess – what would have happened after a patients' withdrawal.

Terminology

Q A lot of current terminology has many different meanings/interpretations (e.g., ITT) do you think this affects the implementation of a shared terminology?

Yes, this makes discussion very difficult. To some extent we need to introduce a new terminology with very specific meaning and allow terms such as ITT, to revert to their original meanings.

Q How do we move to a situation where we have a shared and agreed terminology? What should this terminology look like?

Agreeing what existing terms mean is going to be difficult but not impossible. It is going to require compromise. I suspect many of us are going to have to change the ways we use certain phrases. Indeed some of it may feel wrong to us for some years. To get there we need to get back to the basics and define terms for the underlying events and then build up from there. Most likely we will need to agree a series of word pairs to cover the wide set of scenarios; treatment withdrawal, trial withdrawal and consent withdrawal. Some expressions such as ITT and per protocol, which have multiple interpretations already are best left alone. As such we may need to introduce new terms to describe important concepts. For instance, our collaborative group at the London School of Hygiene and Tropical Medicine have started to use the phrases *de jure* and *de facto* to describe the on-treatment, off-treatment contrast. In terms of understanding missing data, a *de jure* analysis would assume that the subjects remained on treatment and behaved similarly to those in their own treatment arm who continued in the trial. So a classic repeated measures analysis with treatment withdrawal as event is *de jure* in this sense. By contrast, a *de facto* analysis tries to answer what would in fact have happened to the future patient that I discussed at the start of this discussion. This might be represented by a mixed-effects model repeated measures analysis with trial withdrawal as the termination event, or perhaps a retrieved dropout analysis.

Q Do you think that trial analysis methods being transparently stated in write-ups would open discussion on how data are interpreted and get the ball rolling on a uniform approach?

It is often difficult to get journal editors to allow

sufficient space to include proper description of statistical methods. So this will be a difficult battle to win. Also I am surprised how few pharmaceutical statisticians have read the consolidated standards of reporting trials statement. In fact the statement is not proscriptive in the area of missing data, although I would argue it should be. This reflects the fact that its main remit is reporting rather than designing trials. As I have said before, the problem of missing data due to early withdrawal must be addressed from the first stages of trial design. What we need at the reporting stage is a summary of the pattern and type of missing data in each arm along with a description of its potential impact on the results. This will be simplified by a consensus on terminology. We must improve the understanding of medical writers on the importance of missing data and help them to describe the extent and type as well as the impact on study conclusions.

The other reason why the methods used to handle missing data must be included in trial reports is so that any later meta-analyses of those trials can allow for the different approaches used in separate trials.

Financial & competing interests disclosure

J Roger has been an employee of GlaxoSmithKline (GSK) and the content in this article was reviewed internally by GSK and approved. Currently, he holds stocks in GSK and AstraZeneca and has consultancy contracts with GSK and Boehringer Ingelheim. He has no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

References

1 Panel on Handling Missing Data in Clinical Trials, National Research Council. *The Prevention and Treatment of Missing Data in Clinical Trials.* National Academies Press, Washington, DC, USA (2010).

2 European Medicines Agency Committee for Medicinal Products for Human Use (CHMP). *Guidline on Missing Data in Confirmatory Clinical Trials.* EMA/CPMP/EWP/1776/99 Rev. 1 (2010).