Published in final edited form as:

Curr Med Res Opin. 2008 July ; 24(7): 2001–2008. doi:10.1185/03007990802215315.

Approaches to reporting long-term data

Kim A. Papp^{1,*}, Philippe Fonjallaz², Florence Casset-Semanaz², James G. Krueger³, and Knut M. Wittkowski³

Kim A. Papp: kapapp@probitymedical.com; Philippe Fonjallaz: Philippe.Fonjallaz@merckserono.net; Florence Casset-Semanaz: Florence.Casset-Semanaz@merckserono.net; James G. Krueger: jgk@mail.rockefeller.edu; Knut M. Wittkowski: kmw@mail.rockefeller.edu

¹ Probit Medical Research, Waterloo, ON, Canada ² Merck Serono International S.A., Geneva, Switzerland ³ The Centre for Clinical Human Genetics Translational Science at the Rockefeller University, New York, NY, USA

Abstract

Background—Long-term clinical studies are essential for monitoring the effectiveness and safety of a drug. Information provided by long-term clinical studies complements the results of short-term, randomized, controlled trials, which often form the basis of regulatory approval for a new drug application. However, with increasing study duration, the use of placebo becomes less ethical, forcing an open-label study design where a reference estimate for comparison, a placebo-treated cohort, is no longer available. Moreover, as the duration of a study increases and the number of patients continuing in the study declines, missing data become more of a problem: they may bias the results. Therefore, standard analytical strategies used in short-term randomized, controlled trials (intent-to-treat, per-protocol) may not always be appropriate for data generated in long-term studies.

Discussion—We suggest using an intent-to-observe population in long-term studies, applying at least three different methods for handling missing data, testing for bias as a sensitivity analysis and reporting results of more than one method if they differ from one another. The use of multiple analyses is supported by regulatory authority and expert guidelines, although it has not been widely adopted in the medical literature.

Summary—Given the inherent limitations of accounting for missing data with each method, the multiple-analysis approach provides more information with which to make better informed decisions, and clearly defined multiple analytical methods may prevent misleading conclusions from being drawn.

Background

Various prospectively defined statistical methods can be used to extract meaningful information from clinical data. Double-blind, randomized, placebo-controlled trials (RCTs)

*Corresponding author: Dr. Kim A. Papp, Probit Medical Research, Waterloo, ON, N2J 1C4 Canada; Tel: 519 579 9535; Fax: 519 579 8312.

Contributors

All authors contributed to the writing and revising of the manuscript and approved the final version.

Competing interests

Philippe Fonjallaz and Florence Casset-Semanaz are both employees of Merck Serono International S.A. The preparation of this manuscript was supported financially by Merck Serono International S.A.

Additional files provided with this submission:

Additional file 1: d2 methodology review references.enl, 68K
<http://www.biomedcentral.com/imedia/3190408191352857/supp1.enl>

have been universally adopted as the standard approach for measuring the short-term clinical efficacy and safety of a drug. The overall objective of a clinical trial is to provide a valid prospective assessment of the difference between treatments with respect to a clinically relevant outcome.

Although the information provided by RCTs allows the suitability of new treatments to be evaluated before making them widely available to patients, long-term studies are needed for monitoring effectiveness and long-term safety, particularly for studies in patients who require chronic treatment for their disease. However, as the duration of a study increases, missing data become more of a problem and can introduce bias in the results [1]. Furthermore, as study duration increases, the use of placebo becomes less ethical and is often not accepted. Therefore, the methodology used to analyze data from short-term studies may not always be appropriate for data generated in long-term studies. The use of sensitivity analysis becomes more important to ensure robustness of the results.

With the increasing use of evidence-based medicine, continual education and vigilance are required to ensure the veracity and applicability of data. We first outline the established analytical approaches currently used in short-term RCTs. We then provide a review of the analytical issues associated with missing data, issues particularly relevant in long-term studies. Methods commonly employed to handle missing data in studies involving categorical efficacy data will be discussed. Examples of how these methods may affect the study outcome are presented. Finally, we outline an analytical approach that we believe appropriate for long-term trials.

Discussion

Analytical approaches used in short-term RCTs

On completion of a short-term RCT, two data sets (also referred to as populations) are used for statistical analyses: the intention-to-treat (ITT) and per-protocol (PP) populations (Table 1). For most studies, these two populations provide similar results. The ITT analysis is presented most often.

ITT analysis—The ITT population is the standard primary analysis set used in clinical trials. This standard population has been defined as a set that “includes all randomized patients in the groups to which they were randomly assigned, regardless of their adherence to the entry criteria, regardless of the treatment they actually received, and regardless of subsequent withdrawal from treatment or deviation from the protocol” (e.g. wrong treatment received, patient dropped-out, non-compliance) [2]. The ITT analysis compares the originally randomized treatment assignment arms and considers all patients randomized[3].

PP analysis—The PP population (also known as the ‘adherers-only’ population) includes only those patients who did not deviate from the protocol [4,5]. Analyses of this patient population will reflect the optimal effect of an intervention when taken as recommended. It is worth noting that consensus guidelines by the International Conference on Harmonisation, a collaboration between experts and the regulatory authorities of Europe, Japan and the USA, states that “*it is usually appropriate to conduct both an analysis of the full analysis set [almost always the ITT population] and a per-protocol analysis.*” [4]

Analytical approaches in long-term trials

In almost every study, and for a variety of reasons, data will be missing with reasons including: patient withdrawal from lack of efficacy, side-effects or relocation; or unavailability of data at certain timepoints because measurements were not taken at a particular study visit, because of

a missed visit or because of non-compliance. The likelihood of missing data becomes greater for longer-term trials. Incomplete data can have a considerable effect on study results as the amount, distribution, and reasons for missing data may all introduce bias. The potential for bias is increased further in the absence of a control group. Control groups are rarely included in long-term studies.

Although considerable efforts are made to minimize information loss, it remains a prevalent complication in the analysis of data from clinical studies. Therefore, before completion of a clinical study, consideration must be given to which patients should be included in the final analyses, and how missing data should be handled. Importantly, these statistical methods must be defined *a priori*. The choice of a statistical approach will depend on the therapeutic area, the objective, the endpoint, and the design of each study.

Analysis sets—ITT and PP approaches consider patients according to the randomized arms within a study. However, long-term trials usually have an open-label design, without a parallel, randomized, control cohort. Without a comparator arm, intergroup comparisons cannot be made and observations are reported descriptively. Therefore, ITT and PP approaches are not appropriate for open-label, long-term studies. We propose that an intention-to-observe (ITO) population should be considered as more relevant in long-term studies. An ITO population includes all patients entering the open-label, observational phase of a long-term study. The most appropriate methods for handling missing data, commonly referred to as ‘imputation’ of missing data, would then be selected and applied to this population.

Approaches for handling missing data—For any missing data, a number of approaches can be taken to provide an estimated value for each missing datum. Here, we describe some of the methods used most commonly. The simplest approach is to assign the value as a success (referred to here as ‘missing equals success’ [MES]) or a failure (referred to here as ‘missing equals failure’ [MEF]; also known as non-responder imputation [NRI]). Alternatively, missing values can be excluded entirely from the analysis (referred to here as ‘missing equals excluded’ [MEX], also known as the as-treated approach) [6]. MES and MEF are the extreme estimates, with MES assuming the best-case scenario of response for missing data and MEF assuming the worst-case scenario of response for missing data. MES analysis tends to provide an optimistic estimate of effectiveness, while MEF will provide a pessimistic or conservative estimate.

Except for the extremes, MES and MEF, the potential impact of missing data depends significantly on the mechanisms that lead to the missing data. These mechanisms must be considered when determining an estimated value for a missing data point. There are three types of missing data, or ‘missingness’: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) [7].

Data are classified as MCAR if the missingness does not depend on (is explained by) either the previously or the subsequently observed outcomes [7]. For MCAR missingness, it can be assumed that the proportion of successes among the missing outcomes is the same as that among the observed outcomes.

If missingness is at random, it is assumed that the true value of the missing data may depend upon data observed previously. Therefore, MAR data need to be populated using a method that takes into account previous data. Numerous methods for imputing MAR data have been proposed, examined and implemented. For instance, for each patient’s missed outcome assessment, the probability of a success or a response is the proportion of successes or responders among the observed outcomes (as in MEX), but only among those patients who had the same outcomes as the patient with the missing observation at other timepoints (e.g. at

the previous timepoint). Additional imputation approaches include carrying forward the worst previous observation (worst-case scenario), carrying forward the best previous observation (best-case scenario), last observation carried forward (LOCF), regression analysis [8], and likelihood-based mixed-effects analysis [9]. However, each approach requires assumptions about the mechanism of missing data. The mechanisms of missing data cannot be verified completely in most clinical trials [10].

Although not generally favored by statisticians, LOCF is the most common imputation method for handling missing data in short-term RCTs. The main feature of LOCF is that estimates are not based on a defined model about what causes data to become missing. LOCF is generally considered to provide a conservative estimate of the efficacy of an intervention because in placebo-controlled superiority trials, it frequently favors the null hypothesis: there is no difference between the test intervention and placebo [11,12]. However, the LOCF has implicit limitations and is not always a 'conservative' approach. For example, if many patients drop out when they 'feel good', LOCF tends to give a high estimate of success. Generally, LOCF is not suitable for imputation of missing data in long-term clinical trials.

Data are considered MNAR if missingness depends on the current unobserved outcomes. Generally, if missing data are neither MAR nor MCAR then they are considered to be MNAR. MNAR data are considered to be non-ignorable; this implies that more information may be needed to obtain imputed values. Clinical trials, therefore, seek to minimize the amount of non-ignorable MNAR data.

Unless unobserved values are MCAR, they may lead to loss of between-group comparability and potentially introduce bias into estimations of treatment effect [13]. Missing data, for whatever reason, may lead to an underpowered trial. The magnitude of this problem can be quantified by simply recalculating the study power using the actual number of observations made. Unfortunately, MNAR data lead to biased populations and, consequently, biased analyses. Imbalance and non-comparability may be introduced if the causes of missing data depend on the process causing the deviation: for example, discontinuations and withdrawals because of adverse events are frequently directly associated with treatment. Missing follow-up data due to patient relocation is a mechanism for missing data that is unlikely to be associated with treatment and, therefore, less likely to cause imbalance and non-comparability between treatment groups.

To allow a reader of a reported study to make the most informed decisions possible, reasons for withdrawal and loss of follow-up must be presented. When patients are excluded from analyses, reasons for exclusion should be stated. In addition, guidance to help the reader interpret results from analyses with imputed data would be of value. For example, a comparison of baseline characteristics for observed and unobserved patients may indicate specific subgroups that are more likely to be excluded.

Application of different analytical approaches

The choice of imputation method used to handle missing data can have a considerable effect on the reported results and may influence whether a treatment difference is statistically significant. Below we describe two examples, one hypothetical and one real, that illustrate this point.

Hypothetical model—We have constructed a very basic hypothetical model, with a deliberately small sample size ($n = 20$), to illustrate how reported efficacy outcomes may differ when different methods of imputing data are applied to the same dataset (Figure 1). In reality, it is unlikely that statistically or clinically meaningful conclusions could be drawn from such a small dataset, but this example illustrates the concepts and results of imputing missing data.

The top half of Figure 1 shows the treatment outcome for each patient. Patients are considered as either a ‘responder’ or a ‘non-responder’ after receiving a single intervention, with results shown for each 3-month timepoint in a hypothetical 3-year study. The absence of a rectangle indicates missing data. Figure 1b is a graph of the percentage of responders at each timepoint for each of the methods of imputation. At the final, month-36 timepoint, it can be seen that for the LOCF analysis, 75% (15/20) of patients would be classed as responders. The MEF analysis provides the most ‘conservative’ estimate, with 55% (11/20) of patients classed as responders, and the MES analysis gives the least conservative estimate, with 85% (17/20 patients) classed as responders. The MES imputation is seen to be optimistic at all timepoints, while MEF is consistently pessimistic at all timepoints. Both LOCF and MEX consistently fall between MES and MEF but vary in their relative conservatism.

Example based on clinical data—A study published recently investigated the efficacy of efalizumab (a recombinant, humanized, monoclonal IgG₁ antibody) for up to 27 months in 339 patients with psoriasis. Analyses were performed on each 3-month treatment segment throughout the study, which encompassed a 3-month First Treatment Period followed by a 30-month Maintenance Period. The period between months 34 and 36 constituted an optional transitional period prior to the commercial launch of efalizumab. The LOCF method was used to impute missing data for each 3-month segment in the ITO population after the initial 3-months of the study, but up to month 3, patients were classified as non-responders for the remainder of the trial if they discontinued treatment [14]. The primary efficacy measure was the percentage of patients who achieved an improvement of $\geq 75\%$ in Psoriasis Area Severity Index (PASI) score (known as a PASI-75). The LOCF imputation and analysis gave a more conservative estimate of efficacy at month 27 (47% of patients achieved a PASI-75). A MEX imputation and analysis was also conducted and provided more optimistic results (72% achieved PASI-75; Figure 2) [14].

Although both the LOCF and the MEX analyses were reported, only the LOCF data were presented when the full results were published [14]. Clearly, these two approaches gave quite different results in terms of efficacy, but only one was presented in the peer-reviewed publication, as is often the case. Of course, if patients withdraw primarily because of inefficacy, a MEX approach may be biased, presenting an overestimate of clinical effectiveness. However, as was the case in this study [14], there are many other reasons for dropping out, including side-effects, study continuation eligibility criteria, pregnancy, geographic relocation, and patient treatment preferences. In these situations, a MEX imputation and analysis may be informative and complement other types of imputation and analyses. Moreover, a MEX imputation may reflect more closely the situation in routine clinical practice compared with other approaches; many patients who drop out may not return.

The multiple analysis argument

The type of imputation used to handle missing data may introduce bias. Bias can affect estimation of treatment effect and comparability of treatment groups. Possible bias can be estimated by applying multiple imputation methods for missing data and then testing for the potential bias associated with each by analyzing the variability in results. Presenting the analysis for each imputation method and analyzing the variances is a sensitivity analysis. Because the issue of missing data increases in longer-term trials, the use of multiple analyses is particularly important. Current guidelines support the use of more than one analysis set. The Consolidated Standards of Reporting Trials (CONSORT) guidelines also suggest that for studies where non-compliance is an issue, several analyses should be considered [5].

The Committee for Proprietary Medicinal Products (CPMP) guidelines recommend that sensitivity analyses are performed, and suggest that a sensitivity analysis comparing the

outcomes of the full-set analysis (i.e. MES, MEF or LOCF) and a complete-case analysis (i.e. MEX) would be a suitable way to achieve this [1]. The MEX analysis adds further information about consistency of results and provides useful information about patients who received treatment, regardless of whether they deviated from the protocol. Indeed, in the context of human immunodeficiency virus clinical trials data, Hill and Demasi state that “*to understand the intrinsic potency of the antiretroviral regimen under study, ITT analysis needs to be supplemented by standardised as-treated analyses, excluding withdrawals for toxicity or other reasons.*”[15] Further support for providing multiple dataset analyses is provided by the AVANTI study group, who argue that rather than designating any method as inherently ‘good’ or ‘bad’, researchers should present clinical trial results using a range of analyses [16].

Summary

No single statistical analysis is perfect. All of the methods available for the analysis of clinical data suffer drawbacks and limitations based on assumptions made for analyzing incomplete information. Examination of published long-term clinical trials indicates that data from these studies are being presented in a number of different ways. There is currently no consensus on which approach provides the most meaningful information [1,4]. Missing data can seriously bias estimates of treatment effects when a proportion of patients is lost to follow up, which is often the case in long-term clinical studies [12]. The issue of whether the data are MCAR plays a pivotal role in determining bias and may limit any single approach for imputing missing values.

We recommend using the ITO population in long-term studies, applying at least three different approaches for imputing missing data, with MEF, MES and MEX being the minimum. Testing for bias with a sensitivity analysis and reporting results of each method of imputation is also necessary. By presenting the MEF–MES set, the influence of different methods of handling missing data can be assessed to see if the selected analysis creates bias. Indeed, multiple methods of imputation and sensitivity analysis can equally be applied to short-term RCTs. Given the inherent limitations of accounting for missing data in each dataset, the multiple-analysis approach provides more information with which to make better informed decisions and may prevent misleading conclusions from being drawn [12].

Acknowledgments

We thank Tom Potter and Imogen Horsey for their assistance in the preparation of this manuscript.

Abbreviations

CONSORT	Consolidated Standards of Reporting Trials
CPMP	Committee for Proprietary Medicinal Products
ITO	Intention-to-observe
ITT	Intention-to-treat
LOCF	Last observation carried forward
MAR	Missing at random
MCAR	Missing completely at random
MEF	Missing equals failure
MES	Missing equals success
MEX	Missing equals excluded

MNAR	Missing not at random
NRI	Non-responder imputation
PASI	Psoriasis area and severity index
PP	Per protocol
RCT	Randomized, controlled trials

References

1. Committee for Proprietary Medicinal Products (CPMP). Book Points to consider on missing data. City: The European Agency for the Evaluation of Medicinal Products (EMA); 2001. Points to consider on missing data.
2. Fisher, LD.; Dixon, DO.; Herson, J.; Frankowski, RK.; Hearn, MS.; Peace, KE. Intention-to-treat in clinical trials. In: Peace, KE., editor. Statistical issues in drug research and development. New York: Raven Press; 1990.
3. Frangakis CE, Rubin DB. Addressing the complications of intention-to-treat analysis in the combined presence of all-or-none treatment non-compliance and subsequent missing outcomes. *Biometrika* 1999;86:365–379.
4. Book ICH Topic E9: statistical principles for clinical trials. Note for guidance on statistical principles for clinical trials. City: The European Agency for the Evaluation of Medicinal Products (EMA); 1998. ICH Topic E9. statistical principles for clinical trials. Note for guidance on statistical principles for clinical trials.
5. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, Gotzsche PC, Lang T. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Annals of Internal Medicine* 2001;134:663–694. [PubMed: 11304107]
6. Mazumdar S, Liu KS, Houck PR, Reynolds CF 3rd. Intent-to-treat analysis for longitudinal clinical trials: coping with the challenge of missing values. *J Psychiatr Res* 1999;33:87–95. [PubMed: 10221740]
7. Little, RJA.; Rubin, DB. Statistical analysis with missing data. 2. Chichester: Wiley; 2002.
8. Youk AO, Stone RA, Marsh GM. A method for imputing missing data in longitudinal studies. *Ann Epidemiol* 2004;14:354–361. [PubMed: 15177275]
9. Mallinckrodt CH, Clark SW, Carroll RJ, Molenbergh G. Assessing response profiles from incomplete longitudinal clinical trial data under regulatory considerations. *J Biopharm Stat* 2003;13:179–190. [PubMed: 12729388]
10. Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. *Br Med J* 1999;319:670–674. [PubMed: 10480822]
11. Moyle GJ. Truth, lies, and statistical tests. *AIDS Read* 2003;13:117–118. 123–124, 126. [PubMed: 12728867]
12. Wright CC, Sim J. Intention-to-treat approach to data from randomized controlled trials: a sensitivity analysis. *Journal of Clinical Epidemiology* 2003;56:833–842. [PubMed: 14505767]
13. Rubin DB. Inference and missing data. *Biometrika* 1976;63:581–592.
14. Gottlieb AB, Hamilton T, Caro I, Kwon P, Compton PG, Leonardi CL. Long-term continuous efalizumab therapy in patients with moderate to severe chronic plaque psoriasis: updated results from an ongoing trial. *J Am Acad Dermatol* 2006;54:S154–S163. [PubMed: 16488337]
15. Hill A, Demasi R. Discordant conclusions from HIV clinical trials – an evaluation of efficacy endpoints. *Antiviral Therapy* 2005;10:367–374. [PubMed: 15918328]
16. Aidsmap forum reports. How to interpret research studies.
<http://www.aidsmap.com/en/docs/02F26BC6-6841-4612-8647-79D0C10DD043.asp>

Figure 1.

Illustrative example of four methods of handling missing data applied to a small, hypothetical sample of patients (intention-to-treat population). LOCF = last observation carried forward; MEF = missing equals failure; MES = missing equals success; MEX = missing equals excluded.

Figure 2. PASI-75 (improvement of $\geq 75\%$ in Psoriasis Area Severity Index [PASI] score) responses with 95% confidence intervals for data analyzed using the last observation carried forward (LOCF) approach and the missing equals excluded (MEX) approach. Adapted from Gottlieb et al. [14].

Table 1

Description of different approaches for the analysis of data in clinical studies

Approach	Abbreviation	Description	Equivalent terminology
<i>Populations</i>			
Intention-to-treat	ITT	All randomized patients in the groups to which they were randomly assigned	
Per-protocol	PP	All patients who did not deviate from the protocol	Adherers only
Intention-to-observe	ITO	All patients entering the observational phase of a long-term study	Maintenance ITT
<i>Imputation of missing data</i>			
Missing equals success	MES	Missing values are assigned as a success	
Missing equals failure	MEF	Missing values are assigned as a failure	Non-responder imputation
Missing equals excluded	MEX	Missing values are excluded from the analysis	As-treated
Missing completely at random	MCAR	The missingness of the data does not depend on the previously observed or current unobserved outcomes	
Missing at random	MAR	The missingness of data depends on the previously observed values, but not the current unobserved values	
Missing not at random	MNAR	The missingness of data depends on the current unobserved outcomes	
Last observation carried forward	LOCF	The previous observation is used for the missing value	