Choosing estimands in clinical trials with missing data

Peer-reviewed author version

Mallinckrodt, Craig; MOLENBERGHS, Geert & Rathmann, Suchitrita (2017)
Choosing estimands in clinical trials with missing data. In: PHARMACEUTICAL STATISTICS, 16(1), p. 29-36.

DOI: 10.1002/pst.1765
Handle: http://hdl.handle.net/1942/24017

Choosing Estimands in Clinical Trials with Missing Data

Craig H. Mallinckrodt[1], Suchitrita Rathmann[1], and Geert Molenberghs[2]

1. Lilly Research Labs, Eli Lilly and Co, Indianapolis, IN, USA.
2. I-BioStat, Hasselt University, Diepenbeek, Belgium and I-BioStat, Katholieke Universiteit, Leuven, Belgium

Abstract

Recent research has fostered new guidance on preventing and treating missing data. Consensus exists that clear objectives should be defined along with the causal estimands; trial design and conduct should maximize adherence to the protocol specified interventions; and, a sensible primary analysis should be used along with plausible sensitivity analyses. Two general categories of estimands are: effects of the drug as actually taken (de-facto, effectiveness) and effects of the drug if taken as directed (de-jure, efficacy). Motivated by examples, we argue that no single estimand is likely to meet the needs of all stakeholders, and that each estimand has strengths and limitations. Therefore stakeholder input should be part of an iterative study development process that includes choosing estimands that are consistent with trial objectives. To this end, an example is used to illustrate the benefit from assessing multiple estimands in the same study. A second example illustrates that maximizing adherence reduces sensitivity to missing data assumptions for de-jure estimands, but may reduce generalizability of results for de-facto estimands if efforts to maximize adherence in the trial are not feasible in clinical practice. A third example illustrates that whether or not data after initiation of rescue medication should be included in the primary analysis depends on the estimand to be tested and the clinical setting. We further discuss the sample size and total exposure to placebo implications of including post-rescue data in the primary analysis.

Key Words: Missing Data, Clinical Trials, Estimands

1. Introduction

Missing data is an incessant problem in clinical trials that can bias treatment group comparisons and inflate rates of false negative and false positive results (1-8). Fortunately, missing data has been an active area of investigation with many advances in statistical theory and in our ability to implement that theory (1, 3, 5, 7, 8). This research set the stage for new and updated guidance for preventing and handling missing data in clinical trials. Most notably, an expert panel from the National Research Council (NRC) that was commissioned by FDA issued an extensive set of recommendations (3).

The NRC recommendations set forth an overarching framework for tackling the problem of missing data. Key pillars of that framework include: 1) clear specification of trial objectives, including defining the causal estimands; 2) trial design and conduct to maximize adherence to protocol defined interventions; 3) and, a sensible primary analysis supported by plausible sensitivity analyses that assess robustness of results to assumptions about the missing data.

The need for clarity in estimands is driven by ambiguities that can arise from missing data. Data may be intermittently missing or missing due to dropout. Patients may or may not be given rescue medications. Assessments after withdrawal from the initially randomized medication or after the addition of rescue medications may or may not be taken, and if taken, may or may not be included in the analyses depending on the estimand being evaluated (9).

Conceptually, an estimand is simply the true population quantity of interest (3); this is specific to a particular parameter, time point, and population. Given the emphasis on clear objectives, the

choice of the primary estimand has been the subject of considerable discussion (1, 3, 4, 8-17). With the variety of clinical trial scenarios and missing data possibilities, consensus on a universally best estimand is neither realistic nor desirable. Therefore, attention has shifted to how to choose estimands.

For clarity the following distinction is made. *Patient dropout* occurs when the patient discontinues the initially randomized study medication and no further observations are taken. *Analysis dropout* occurs when patients deviate from the originally randomized treatment regime (stops medication and/or adds rescue medication) and observations are taken but they are not included in the analysis because they are not relevant for the estimand being evaluated.

Leuchs et al (11) proposed a process chart that begins with defining the primary estimand, followed by design, analysis, and sensitivity analyses. In a letter to the editor in response to the Leuchs et al. paper the PSI/EPSI working group (WG) on estimands advocated an additional step prior to choosing the primary estimand. The WG proposal began by considering objectives and then proceeded to the other steps in an iterative manner so that interactions between the various components could be considered (12). The WG expanded on these ideas in a subsequent paper (17).

The purpose of the present paper is to utilize this iterative study development process to illustrate key issues in choosing estimands. The paper is organized by providing an overview of objectives and estimands in section 2. Fundamental considerations for estimands are discussed along with considerations for use of rescue data, trial design, and analysis in section 3. These

considerations are illustrated via examples in section 4. Points are then tied together in the discussion in section 5.

2. Objectives and Estimands

Trial objectives are typically driven by the decisions to be made from the trial results. These decisions depend in part on stage of development. Phase II trials are typically used by drug developers to determine proof of concept or to choose doses for subsequent studies. Phase III, confirmatory studies typically serve a diverse audience and address diverse objectives (11). For example, regulators render decisions regarding whether or not the drug under study should be granted a marketing authorization. Drug developers and regulators must collaborate to develop labeling language that accurately and clearly describes the risks and benefits of approved drugs. Payers must decide if / where a new drug belongs on its formulary list. Prescribers must decide if a drug should be prescribed to particular patients and must inform those patients and / or their care givers of what to expect. Patients and care givers must decide if they want to take the drug that has been prescribed.

With a clearly defined objective and estimand, appropriate analytic choices become clearer. Estimands can be divided into two general categories, efficacy and effectiveness. Efficacy may be viewed as the effects of the drug if taken as directed. Effectiveness may be viewed as the effects of the drug as actually taken (1, 6, 10, 11, 14, 15). However, the efficacy and effectiveness nomenclature does not make sense for safety outcomes. A more general terminology is de-jure (if taken as directed) and de-facto (as actually taken) (13).

The NRC guidance (3) discusses five estimands in detail and Mallinckrodt et al (10) proposed a 6th estimand. The focus here is on three of those estimands to illustrate many of the considerations in choosing estimands. Each of the three estimands involves the difference versus control in changes to the planned endpoint of the trial, in all randomized patients.

1. Estimand 1 is the change due to the treatment regimens as actually taken.
2. Estimand 2 is the change due to the initially randomized treatments as actually taken.
3. Estimand 3 is the change due to the initially randomized treatments if taken as directed.

The distinction between estimands 1 and 2 is whether or not data after discontinuation of the initially randomized treatment and / or initiation of rescue treatment are needed to estimate the estimand. Post-discontinuation and post-rescue data are needed for estimand 1, but are not needed for estimand 2.

It is also important to differentiate estimand 3, which is based on all randomized patients, from what is estimated in a completers analysis. In a completers analysis the estimand being evaluated is conditional on having been adherent. Completers analyses do not preserve the initial randomization. In contrast, estimand 3 includes all randomized patients. Therefore, in principle, inferences and parameter estimates from estimand 3 apply to all patients in the population, not merely to those who were doing well enough to remain adherent.

Another potential de-jure estimand not discussed in detail here is the de-jure estimand in patients who were shown to tolerate the experimental drug during a run-in phase. This estimand was discussed in the NRC guidance (3). The intent of this “tolerator” estimand is similar to that of estimand 3. However, the tolerator estimand draws inference regarding drug benefit from the

subset of patients that tolerated the drug during the run-in phase and were subsequently randomized. Safety assessments would have to come from all patients exposed to the experimental drug. Therefore, when this design is used inference for safety and efficacy cannot be drawn from the same group of patients.

The following table relates the numbering of the 3 focus estimands in this paper to the numbering used in the NRC guidance (3) and in Mallinckrodt et al (10).

Table 1. Number references of estimands in various publications

	Reference number	
This paper	NRC guidance	Mallinckrodt et al (10)
1	1	1
2	estimand not mentioned	6
3	3	3

3. Considerations

Fundamental considerations

Given the diversity in clinical settings and decisions to be made from clinical trial data, no universally best primary estimand exists, and therefore multiple estimands are likely to be of interest for any one trial (1, 10, 11, 12, 14, 15, 17).

The three common estimands defined in Section 2 each have strengths and limitations. The de-jure estimand (estimand 3) can be considered hypothetical (i.e., counterfactual) for groups of patients because treatment effects are assessed as if taken as directed when in any meaningfully

sized group some patients will not adhere (3). However, the de-jure estimand is relevant because knowing what to expect when patients adhere is important. Patients are advised to take their medication as directed; therefore, it is important to assess what happens if a medication is taken as directed so that optimal directions can be developed. Example 1 in Section 4 illustrates the role estimand 3 can play in optimizing drug development and patient care..

De-facto estimands can be considered counterfactual for individual patients because treatment effects are assessed from a mix of adherent and non-adherent patients, but each patient is either adherent or not adherent, no patient is both. On the other hand, de-facto estimands can provide useful estimates of what to expect from the group as a whole (2, 3).

Most of the discussion on de-jure versus de-facto estimands has been in the context of assessing drug benefit. However, estimands for assessing drug risk are also important. Consider the following hypothetical example. A drug has the adverse effect of increasing blood pressure. Some patients become hypertensive and discontinue study medication and/or take rescue medication, with subsequent return to normal blood pressure. De-facto estimands would reflect the patients' return to normal, thereby suggesting no change in blood pressure at the planned endpoint of the trial. De-jure estimands would not reflect a return to normal and would reflect increases at endpoint because had the patients been adherent they would likely have continued to be hypertensive. Therefore, for safety assessments de-jure estimands may be particularly relevant.

Another important fundamental consideration is whether or not the estimand is consistent with the intention-to-treat (ITT) principle. The ICH E9 guidance (18) defines ITT as "the principle that asserts that the effect of a treatment policy can be best assessed by evaluating on the basis of the intention to treat a subject (i.e. the planned treatment regimen) rather than the actual treatment given. The guidance parses ITT into two parts, the patients to include and the data for each patient to include. The guidance is clear on the need to include all randomized patients.

The de-jure estimand based on completers is clearly not consistent with ITT because it does not include all randomized patients. In contrast, consider de-jure estimand 3 defined in Section 2. This estimand includes all randomized patients and is therefore consistent with ITT in that regard. Additional aspects of ITT are covered in the next section. Interestingly, an FDA working group on missing data noted common situations in which estimands could be adequately addressed by patient subsets that did not preserve the initial randomization (16).

Rescue medication considerations

Whether or not data collected after discontinuation of initially randomized study medication or initiation or rescue medication should be included in the primary analysis is an important consideration. Rescue medication can mask or exaggerate the (efficacy and safety) effects of the initially assigned treatments, thereby biasing estimates of the effects of the originally assigned medication (1, 10, 14, 15, 20, 21, 22).

The ICH E9 guidance (18) definition of ITT states ”that subjects allocated to a treatment group should be followed up, assessed and analyzed as members of that group irrespective of their compliance to the planned course of treatment.” Importantly, the E9 guidance refers to ITT in the context of assessing treatment regimens and does not address inference for the initially randomized medications, such as is the case with estimands 2 and 3.

Rescue therapy is specifically addressed in ICH E10, section 2.1.5.2.2 (19). In referring to trials with rescue, E10 states: “In such cases, the need to change treatment becomes a study endpoint.” This seems to suggest that at the point of rescue it is known the treatment did not work for that patient and that post-rescue data need not be included in the primary analysis. Instead, the primary analysis might include adherence and need for rescue as part of a composite endpoint or as the sole primary endpoint. However, the estimand is not explicitly stated in the E10 guidance and therefore ambiguity remains about if and when post-rescue data should be included in the primary analysis. The ICH guidance was issued prior to the more nuanced discussion of estimands that is taking place today; hence, the need for updated ICH guidance on estimands. An addendum to the E9 guidance on estimands and sensitivity is anticipated in 2016.

The perceived need for rescue therapy may be partly motivated by arguments for ethical patient care, especially in placebo-controlled trials. However, in placebo controlled trials if rescue therapy is beneficial including post-rescue data is likely to substantially decrease the magnitude of the treatment effect compared with de-jure or de-facto estimands that exclude post-rescue data. Therefore, to maintain power an increased sample size is needed, which exposes more patients to placebo (20).

With estimand 1 (de-facto, treatment regimens) data after discontinuation of the initially randomized medication and/or addition of rescue medication are included in the analyses (3). Therefore, inference for estimand 1 is in regards to treatment regimens (1, 7, 9, 10, 11, 14, 15). However, the most relevant questions in early research and initial regulatory review, especially for placebo-controlled trials, are often about the effects of the investigational drugs, not treatment regimens involving the investigation drug (1, 10, 15). The very fact the post-rescue data are not often collected and / or included in the primary analysis of confirmatory trials in many settings suggests the pragmatic effectiveness estimand is not the one of primary interest (16).

O'Neill and Temple (14) noted that primary estimands requiring data after withdrawal of randomized medication and / or initiation of rescue may be more common in outcomes trials where the presence / absence of a major health event is the endpoint and/or the intervention is intended to modify the disease process. Symptomatic trials (symptom severity is the endpoint) typically focus on inferences regarding the initially randomized treatments. Symptomatic trials can avoid the confounding from rescue medications by using a primary estimand and analysis that exclude data after discontinuation of study medication / initiation of rescue.

Conceptually, estimand 2 (de-facto, initially randomized treatments) avoids the confounding effects of rescue medications on inferences regarding the initially randomized treatments by not allowing rescue medication. However, given the ethical mandate to allow rescue medications

and the analytic need to exclude post-rescue data the issue of how to estimate estimand 2 is covered in the subsection on analysis considerations.

For estimand 3 (de-jure, initially randomized treatments) data after discontinuation of treatment or initiation of rescue are not required.

Design considerations

Universal agreement exists that trials should aim to maximize adherence to protocol procedures, including adherence to the initially assigned treatments (1, 2, 3, 8, 14, 15). Maximizing adherence improves robustness of results by reducing the reliance of inferences on the untestable assumptions about the missing data (1, 3, 5, 7, 8, 14, 15, 23). These considerations have often been in the context of de-jure estimands. However, the impact of maximizing retention on de-facto estimands should also be considered (11).

Increasing adherence is likely to increase benefit from the drug as actually taken, thereby resulting in more favorable estimates of de-facto estimands. If the measures used to engender adherence in the clinical trial are not feasible in clinical practice the trial could yield biased estimates of effectiveness relative to the conditions under which the drug would be used.

Specifically, assessment of de-facto estimands often entails using adherence as part of the primary outcome. For example, patients that discontinue study drug are often considered a treatment failure regardless of the observed outcomes. Therefore, it is important to consider the degree to which treatment adherence decisions in the clinical trial match adherence decisions in

clinical practice. These generalizability considerations may be especially important in trials with placebo and / or blinding because these factors are never present in clinical practice (1).

Analysis considerations

In the iterative study development process advocated by Leuchs et al. (11) and the PSI/EPSI working group on estimands (17), it is possible for analytic considerations to influence choice of estimands. For example, interest may center on a particular estimand, but if a robust analysis (and / or design) cannot be paired with this estimand, it may be better to focus on a related estimand for which a robust analysis exists. Therefore, it is important to understand the attributes of analytic approaches when choosing estimands.

The primary analysis of the de-facto treatment regimen estimand (estimand 1) and the de-jure estimand (estimand 3) introduced in Section 2 typically involves the assumption that data are missing at random. This assumption is far less restrictive than missing completely at random, and is at least a good starting point in clinical trial analyses (1, 3, 5, 7, 8, 14, 15, 23). The plausibility of MAR hinges in part on having minimal loss to follow-up. When patients are lost to follow up data explaining why they discontinued study medication are not fully available, making model assumptions more suspect than if such data were available. Regardless, validity of MAR can never be proven; hence, the need for plausible sensitivity analyses (1, 14, 15, 23, 24).

One approach to assessing estimand 2 is to impute the data after initiation of rescue and/or discontinuation of the initially randomized medication under the assumption that initially

randomized medications have no (or diminished) effect after discontinuation / rescue (1, 7, 8, 10, 13, 14, 15). This assumption is often reasonable in trials of symptomatic interventions (4, 25).

Such imputations for continuous endpoints have historically been done using baseline observation carried forward (BOCF). For categorical endpoints non-responder imputation (NRI) has often been used. With NRI all patients that discontinue initially randomized medication and/or initiate rescue medication are considered non-responders, regardless of the outcome observed at the planned endpoint of the trial. However, single imputation approaches such as BOCF and NRI have a number of disadvantages and more principled approaches are gaining favor (1, 10, 13, 15, 26).

In BOCF and NRI, the assumption of no change from baseline is made in order to ascribe no pharmacologic benefit from the drug if it is discontinued. However, these approaches ignore the potential changes from non-pharmacologic sources that are often seen in trials (study effect, placebo effect) and would therefore be valid only in those situations where there was no change in a placebo group over time (1, 10, 15).

The bias in BOCF estimates can be large, resulting in inflated type I error rates or loss of power in testing de-facto estimands (28). In addition, BOCF makes no sense in situations where the therapeutic aim is to prevent worsening because carrying the baseline observation forward ascribes a good outcome to patients that discontinue (1). Moreover, as a single imputation technique, BOCF assigns the same change score (zero) to every patient that discontinues, which

results in underestimates of variance and standard error. Therefore, BOCF is generally not a useful analytic approach (1, 3, 5, 7, 8, 10, 15, 16, 24).

Multiple imputation-based approaches to test de-facto estimands have come into the literature recently. These methods have been referred to as controlled imputation or more specifically reference-based controlled imputation (1, 7, 8, 13, 14, 15). Full descriptions of these approaches go beyond the present scope. However, the general approach is to use multiple imputation in a manner that accounts for the change in / discontinuation of treatment. In so doing, patients that discontinue from an experimental arm have values imputed as if they were in the reference (e.g., placebo arm). Depending on the exact implementation, imputed values can either reflect no pharmacologic benefit from the drug immediately upon discontinuation / rescue, a decaying benefit after discontinuation / rescue, or a constant benefit after discontinuation / rescue (7, 8, 13, 14, 15).

In contrast to BOCF, reference-based imputation via multiple imputation accounts for the uncertainty of imputation, accounts for study / placebo effects and can therefore be applied regardless of whether the therapeutic aim is improvement or prevention of worsening (1, 7, 8, 13, 14). Reference based imputation has been shown to reduce bias and provide better control of type 1 error compared with BOCF (27). Some of these methods can now be implemented in commercially available software (28) and specialty programs have also been made freely available to the public at www.missingdata.org.uk.

4. Examples

Three short examples are used. The first example illustrates the benefits of assessing both de-jure and de-facto estimands in the same trial. The second example illustrates the impact of increasing adherence on de-jure and de-facto estimands regarding the initially randomized treatments. The third example illustrates including post-rescue data in analyses.

First, consider the following hypothetical example where effectiveness is a function of efficacy and adherence. Drug A and Drug B (or dose A and dose B of a drug) have equal effectiveness but A has significantly greater efficacy and B has significantly greater adherence.

	Efficacy	Adherence	Effectiveness
Drug A	High	Low	Average
Drug B	Low	High	Average

These differences in clinical profiles have important implications. Dose / Drug A might be the best choice for patients with more severe illness because it has greater efficacy. Dose / Drug B might be best for patients with less severe illness and / or safety and tolerability concerns because it has greater adherence resulting from fewer side effects. In the context of two doses of a drug the more nuanced understanding of efficacy and adherence could lead to additional investigation that could lead to more optimized patient outcomes. For example, subgroups of patients who especially benefit from or tolerate the high dose might be identified from the existing data or from a new trial (non-responders to low dose). Or, alternate dosing regimens that might improve the safety / tolerability of the high dose, such as titration, flexible, or split dosing (40 mg every two weeks rather than 80 mg every 4 weeks), could be investigated in subsequent trials.

Example 2 is from clinical trials in depression, which have been used in a previous examination of sensitivity analyses (14). These data sets were somewhat contrived to avoid implications for marketed products, but key features of the original data were preserved. The original data were from antidepressant clinical trials (29, 30). Assessments on the Hamilton 17-item rating scale for depression (HAMD17) (31) were taken at baseline and weeks 1, 2, 4, 6, and 8 in each trial. These trials are referred to as the low and high dropout datasets. In the high dropout data set completion rates were 70% for drug and 60% for placebo. In the low dropout dataset completion rates were 92% in both arms.

The design differences that may explain the difference in dropout rates between these two otherwise similar trials was that the low dropout dataset came from a study conducted in Eastern Europe that included a 6-month extension treatment period after the 8-week acute treatment phase, and used titration dosing. The high dropout data set came from a study conducted in the US that did not have the extension treatment period and used fixed dosing.

Estimates of de-facto and de-jure estimands regarding the initially randomized treatments (estimands 2 and 3 in Section 2, respectively) were obtained from each data set in order to illustrate the impact of higher and lower adherence.

The de-facto estimand was assessed by defining each patient as a treatment success or failure. Treatment success was defined as improvement greater than or equal to 50% of the baseline severity and completion of the acute treatment phase. Any patient that discontinued study medication was considered a treatment failure regardless of outcome. Treatment groups were

compared using Fisher's Exact test. The treatment success / failure approach results in the same numeric quantity as non-responder imputation (NRI). However, the result is interpreted differently. The definition of the treatment success included dropout, so there was no missing data and no need to assess sensitivity. In contrast, with NRI, as the name implies, response status is imputed and hence sensitivity would need to be assessed.

The de-jure estimand was estimated using a restricted maximum likelihood (REML)-based repeated measures approach. The analyses included the fixed, categorical effects of treatment, investigative site, visit, the continuous, fixed covariate of baseline score and all two-way interactions with visit. An unstructured (co)variance structure shared across treatment groups was used to model the within-patient errors. The Kenward-Roger approximation was used to estimate denominator degrees of freedom and adjust standard errors. Analyses were implemented using SAS PROC MIXED (30). The primary comparison was the contrast (difference in LSMEANS) between treatments at the last Visit (Week-8).

Sensitivity of the de-jure results to departures from MAR was assessed using the reference-based imputation approach known as jump to references (J2R) (13, 14). In J2R, values for reference group patients are imputed assuming MAR; values for drug treated patients are imputed assuming MNAR with the benefit from the drug immediately disappearing after discontinuation of study drug. The J2R imputations were implemented using the placebo group as the reference group, with a full multivariate repeated measures model for parameter estimation that included treatment, investigative site and baseline score, all crossed with visit. The analysis model was ANOVA at week 8 with treatment, baseline and pooled investigator in the model.

Although J2R can be implemented as an assessment of estimand 2, it can also be implemented as it is here, a worst reasonable case departure from MAR. That is, the same numeric quantity, the estimate from J2R, can be interpreted as either an assessment of estimand 2 or as a sensitivity analysis for estimand 3. In the sensitivity context, J2R assumes an MAR mechanism for the reference arm and an MNAR mechanism for the experimental arm such that the estimate of the treatment effect will be smaller than if MAR was assumed for the experimental arm. A full discussion of sensitivity analyses is beyond the present scope. However, it is important to appreciate that the reference distribution in J2R is not based on only completers; MAR is assumed for the reference arm such that patients with poor outcomes who drop out early contribute. However, as in other applications, the assumption of MAR or any specific MNAR mechanism cannot be validated from the observed outcomes and / or reasons for discontinuation.

Results from example 2 are summarized in Table 2. For both data sets the MAR analysis of the de-jure estimand yielded a significant treatment contrast. In the low dropout dataset the J2R sensitivity analysis yielded a treatment effect very close to the MAR estimate. However, in the high dropout dataset the difference between the MAR and J2R result was 5-fold greater than in the low dropout dataset and statistical significance was not preserved. Therefore, with low dropout inference regarding the de-jure estimand was robust to plausible departure from MAR whereas results from the high dropout dataset were not robust to plausible departure from MAR.

Table 2. Results from Example 2.

	High dropout				Low dropout			
	LSMEANS		LSMEAN		LSMEANS		LSMEAN	
	Placebo	Drug	Difference[1]	Pvalue	Placebo	Drug	Difference1	Pvalue
MAR	-5.95	-8.24	2.29 (1.00)	0.024	-10.56	-12.40	1.84 (0.70)	0.009
J2R	-5.97	-7.57	1.60 (0.99)	0.110	-10.55	-12.26	1.72 (0.70)	0.016
	Treatment success (%)				Treatment success (%)			
	Placebo	Drug	Difference	Pvalue	Placebo	Drug	Difference	Pvalue
	25	39	14	0.034	50	68	18	0.001

1. LSMEANS are mean change from baseline in HAMD17 total score. LSMEAN difference is the contrast between drug and placebo at the planned endpoint assessment (Week 8). Values in parenthesis are the standard errors of the LSMEAN differences.

In regards to the de-facto estimand, both trials yielded significant differences in treatment success. However, generalizability must be considered. The low dropout data set had greater within group mean changes and greater adherence. The percent treatment success on placebo in the low dropout data set was two-fold greater than in the high dropout dataset. It is not certain from this example if or how the design differences influenced the within group mean changes and adherence. However, the two trials did give different views of effectiveness and in a real scenario it would be important to justify generalizability of effectiveness results.

Example 3 is from a clinical trial in psoriatic arthritis. For full details of the study see NCT01695239 in www.clinicaltrials.gov. Patients were randomized in a 1:1:1:1 ratio to two doses of an experimental drug, a known effective standard of care (adalimumab, humira), and placebo. This investigation focuses on only the standard of care and placebo arms. Assessments were taken at baseline, weeks 1, 2, 4, 8, 12, 16, 20 and 24. After week 16, patients with inadequate response to the standard of care or placebo were allowed to have changes in

background therapy as rescue. As additional rescue therapy, patients initially randomized to placebo were re-randomized to one of the two doses of the experimental medication until week 24, the time of the primary assessment. This additional rescue intervention could not be implemented for the standard of care arm due to safety concerns related to switching immediately from one active medication to another.

These data are used to compare results when including versus not including post-rescue data in the analyses. Specifically, this re-analysis compared results from estimand 1 (treatment regimens as actually taken) and estimand 2 (initially randomized treatments as actually taken) with regard to the percentage of patients meeting ACR20 criteria. This variable is a common choice for the primary analysis and essentially assesses whether or not patients had a 20% improvement from baseline in signs and symptoms of their psoriatic arthritis.

For estimand 1 patients were considered a treatment success if they met ACR20 criteria at week 24 and patients were considered a treatment failure if they did not meet ACR20 criteria at week 24, or they discontinued study medication. For estimand 2, patients were considered a treatment success if they met ACR20 criteria at week 24 and did not receive rescue treatment. Patients were considered a treatment failure if they did not meet ACR20 criteria at week 24, or they discontinued study medication, or they required rescue medication.

The difference between the two results reflects the effects of rescue medication. In the adalimumab arm 12 of 101 patients met criteria for rescue treatment and 3 of the patients that took rescue medication met ACR20 criteria at week 24. In the placebo arm, 45 of 106 patients

met criteria for rescue treatment and 13 of the patients that took rescue medication met ACR20 criteria at week 24. Results are summarized in Table 3.

Table 3. Number and percent of patients meeting treatment success criteria with and without inclusion of post rescue data in example 3.

	Adalimumab (n = 101)	Placebo (n = 106)	Difference
Without post rescue data (estimand 2)	58 (57.4%)	32 (30.2%)	27.2 %
With post-rescue data (estimand 1)	61 (60.4%)	45 (42.5%)	17.9 %

Treatment success rates without post-rescue data showed an advantage of drug over placebo of approximately 27% versus an advantage of 18% when post-rescue data were included. Treatment groups differed significantly ($p < .01$) regardless of whether or not post-rescue data were included. However, if the rate of success among patients eligible for rescue in the placebo arm had been 17 of 45 (38%) instead of 13 of 45 (29%), significance would have been lost when including post-rescue data

Powering a future study based on results of estimand 2 suggest 75 patients per arm would yield 90% power, whereas results from estimand 1 suggest 175 patients per arm are needed for 90% power, thereby more than doubling the exposure to placebo.

5. Discussion

Consensus exists that the best way to deal with missing data is to prevent it and that a sensible primary analysis should be supported by plausible sensitivity analyses that assess robustness of inferences to violations of missing data assumptions. Consensus also exists on the need for clarity in objectives and estimands. However, given the diverse settings in which clinical trials are conducted it is neither realistic nor desirable to seek consensus on a universally best primary estimand, primary analysis, or approach to sensitivity analyses. Therefore, discussion is shifting toward the process by which these decisions are made.

Leuchs et al (11) suggested a process chart that begins with choice of the primary estimand, after which design, analysis and sensitivity analyses could be determined. The PSI / EFSPI WG advocated a refinement of that proposal wherein trial objectives drive choice of estimands in an iterative process to allow design and analysis considerations to be factored into the choice of estimand(s) (12, 17).

The intent of the present paper is not to illustrate specific choices of estimands for specific situations. Rather, the intent is to examine considerations for choosing estimands using three focus estimands. These three are not the only estimands of interest in clinical trials. For example, an estimand noted in the NRC guidance (3) that was not discussed here is the de-jure estimand in patients who were shown to tolerate the experimental drug during a run-in phase.

Examples were used to illustrate the benefits of multiple estimands in the same trial, the consequences of increasing adherence, and the consequences of including post-rescue data. In

the iterative process proposed by the WG, design and analysis considerations can influence objectives and estimands. This is not in conflict with the notion that handling of missing data should not compromise the meaningfulness of endpoints and estimands (32). Rather, jointly considering all aspects of the process can lead to objectives, estimands and designs that are more relevant given the circumstances.

For example, consider a six-week acute phase clinical trial where it is anticipated that extensive efforts to maximize adherence will yield 95% of patients remaining on the initially assigned study medication. With this level of adherence plausible departures from MAR are unlikely to overturn positive findings for a de-jure estimand. A de-jure primary estimand is consistent with this highly controlled setting that yields high levels of adherence. However, when assessing long–term effects in the same disease state, the rigid control of the short term trial may be too burdensome on patients. A less restrictive design that is more similar to clinical practice may be needed. A de-facto primary estimand is consistent with the more pragmatic nature of the trial and with the inevitable loss of adherence over the longer treatment period.

As another example of interplay between estimands, design, and analyses consider example 3 in the previous section. The primary assessment time was week 24, with rescue first available at week 16. The more nuanced discussions of estimands that are present today compared with when the example study was planned may have led investigators to choose a primary endpoint of week 16 if estimand 2 or estimand 3 was primary, but week 24 if estimand 1 was primary. If estimand 1 was chosen as primary, that choice may have influenced choice of comparator. Estimand 1 is often used as a pragmatic assessment of effectiveness (16). However, placebo is

never used in clinical practice. Therefore, placebo control may be less consistent with estimand 1 than active control.

As another example of how other factors can influence choice of estimand, consider a trial where focus is on effectiveness, but interest is in both estimand 1 and estimand 2; that is, results with and without post-rescue data are relevant. Also consider that in this scenario it is important to keep the sample size as small as possible either because patients are hard to recruit, or for ethical reasons it is important to limit exposure to placebo. Use of estimand 2 as the primary estimand is likely to result in greater power, which translates into smaller sample sizes and reduced exposure to placebo. Post-rescue data may still be collected and used secondarily.

The multifaceted nature of clinical trials is important to consider in choosing estimands (10, 11, 12, 17). Objectives early in development often differ from objectives later in development. Even within confirmatory trials, diverse objectives are needed to inform the decisions regulators, health technology assessors / payers, prescribers, patients, caregivers, sponsors, other researchers, etc. must make. Even for a single stake-holder in a single trial it is often important to know what happens when a drug is taken as directed (de-jure estimand) and to know what happens when the drug is taken as in actual practice (de-facto estimand). Therefore, no single estimand is likely to best serve the interests of all stake holders and de-jure and de-facto estimands will both be of interest (1, 10, 11, 12, 17).

By including de-facto and de-jure estimands in a single trial those that make decisions about individual patients (prescribers, patients, caregivers, etc.) may focus most on the de-jure

estimands and secondarily on de-facto estimands. Those that make decisions about groups of patients (e.g., regulators, HTAs) may focus most on de-facto estimands. However, all decisions makers will benefit from understanding results from both de-jure and de-facto estimands.

Conclusions

An iterative process should be used to choose estimands, beginning with the objectives required to address the needs of diverse stake-holders. No single estimand is likely to meet the needs of all stake-holders. De-jure and de-facto estimands each have strengths and limitations. Fully understanding a drug's effects requires understanding results from both families of estimands. Maximizing adherence reduces sensitivity to missing data assumptions for de-jure estimands. However, it is also important to consider generalizability of results for de-facto estimands if efforts to maximize adherence in the trial are not feasible in clinical practice. Whether or not data after initiation of rescue medication should be included in the primary analysis depends on the estimand to be tested and the clinical setting.

References

1. Mallinckrodt, CH. Preventing and Treating Missing Data in Longitudinal Clinical Trials: A Practical Guide. 2013; Cambridge University Press. New York.

2. Committee for Medicinal Products for Human Use (CHMP). Guideline on missing data in confirmatory clinical trials. 2010. EMA/CPMP/EWP/1776/99 Rev. 1

3. National Research Council (2010). The prevention and Treatment of Missing Data in Clinical Trials. Panel on Handling Missing Data in Clinical Trials. Committee on National Statistics, Division of Behavioural and Social Sciences and Education. Washington, DC: The National Academies Press.

4. O'Neill RT and Temple R. (2012). The Prevention and Treatment of missing Data in Clinical Trials: An FDA Perspective on the Importance of Dealing With It. Clinical Pharmacology and Therapeutics. doi:10.1038/clpt.2011.340

5. Molenberghs G, Kenward MG. (2007), Missing Data in Clinical Studies. Chichester: John Wiley & Sons.

6. Mallinckrodt CH, Lane PW, Schnell D, Peng Y, and Mancuso JP. Recommendations for the Primary Analysis of Continuous Endpoints in Longitudinal Clinical Trials. Drug Information Journal. 2008; 42:305-319.

7. Molenberghs G, Fitzmaurice G, Kenward M, Tsiatis A, Verbeke G. Handbook of Missing Data Methodology. (2015). CRC Press, Boca Raton

8. O'Kelly M and Ratitch B. Clinical Trials with Missing Data. (2014). Wiley, Chichester

9. Mallinckrodt CH, Kenward MG. Conceptual considerations regarding choice of endpoints, hypotheses, and analyses in longitudinal clinical trials. Drug Information Journal. 2009; 43(4):449-458.

10. Mallinckrodt, CH, Lin Q, Lipkovich I, Molenberghs G. A structured approach to choosing estimands and estimators in longitudinal clinical trials. Pharmaceutical Statistics. 2012, 11 456–461

11. Leuchs, AK, Zinserling J, Brandt A, Wirtz D, Benda N. Choosing appropriate estimands in clinical trials. Therapeutic Innovation and Regulatory Science. February 17, 2015, doi: 10.1177/2168479014567317.

12. Garrett A. Choosing Appropriate Estimands in Clinical Trials (Leuchs et al): Letter to the Editor. Therapeutic Innovation & Regulatory Science. May 6, 2015 doi:10.1177/21684790155860.

13. Carpenter J, Roger J, and Kenward M. Analysis of Longitudinal Trials with Missing Data: A Framework for Relevant, Accessible Assumptions, and Inference via Multiple Imputation. 2013. J Bio pharm Stat 23:1352-1371

14. Mallinckrodt CH, Roger J, Chuang-Stein C, et al. Recent Developments in the Prevention and Treatment of Missing Data. Therapeutic Innovation and Regulatory Science; 2014, 48(1): 68-80.

15. Mallinckrodt CH, Roger J, Chuang-Stein C, et al. Missing data: Turning guidance into action. Statistics in Biopharmaceutical Research; 2013, 5(4): 369-382.

16. Permutt T. A taxonomy of estimands for regulatory clinical trials with discontinuations Statis Med. 2015a. DOI: 10.1002/sim.6841

17. Phillips A., Abellan-Andres J., Andersen S. et al. Pharmaceutical Statistics, accepted. 2016. Estimands: Discussion Points from the PSI Estimands and Sensitivity Expert Group. DOI: 10.1002/pst.1745.

18. ICH guidelines. Online at: http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/Step4/E9_Guideline.pdf

19. ICH guidelines. Online at: http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E10/Step4/E10_Guideline.pdf

20. Buncher, C. Ralph and Jia-Yeong Tsay, Statistics in the pharmaceutical industry. 3rd edition. 2005. Chapman Hall/CRC Press.

21. Holubkov R1, Dean JM, Berger J, Anand KJ, Carcillo J, Meert K, Zimmerman J, Newth C, Harrison R, Willson DF, Nicholson C. Is "rescue" therapy ethical in randomized controlled trials? Pediatr Crit Care Med. 2009 Jul;10(4):431-8. doi: 10.1097/PCC.0b013e318198bd13.

22. Henning Zeidler. Paracetamol and the Placebo Effect in Osteoarthritis Trials: A Missing Link? Pain Research and Treatment. (2011). Doi: 10.1155/2011/696791

23. Verbeke G, Molenberghs G. (2000). Linear Mixed Models for Longitudinal Data.

24. Molenberghs, G., Beunckens, C., Sotto, C., and Kenward, M.G. (2008). Every missing not at random model has got a missing at random counterpart with equal fit.

25. Kim Y. Missing Data Handling in Chronic Pain Trials. Journal of Biopharmaceutical Statistics. 2011; 21: 2, 311 – 325.

26. Kenward, M.G. and Molenberghs, G. (2009). Last Observation Carried Forward: A Crystal ball? *Journal of Biopharmaceutical Statistics*, **19**, 872-888.

27. Ayela B, Lipkovich I, Molenberghs G, Mallinckrodt CH. (2014). A Multiple -Imputation-Based Approach to Sensitivity Analyses and Effectiveness Assessments in Longitudinal Clinical Trials, Journal of Biopharmaceutical Statistics, 24:2, 211-228,

28. SAS Institute Inc. 2013. SAS/STAT® 9.4. User's Guide. Cary, NC: SAS Institute Inc.

29. Goldstein DJ, Lu Y, Detke MJ, Wiltse C, Mallinckrodt C, Demitrack MA: Duloxetine in the treatment of depression: a double-blind placebo-controlled comparison with paroxetine. J Clin Psychopharmacol 2004;24: 389-399.

30. Detke MJ, Wiltse CG, Mallinckrodt CH, McNamara RK, Demitrack MA, Bitter I. Duloxetine in the acute and long-term treatment of major depressive disorder: A placebo- and paroxetine-controlled trial. European Neuropsychopharmacology. 2004; 14(6):457-470

31. Hamilton M: A rating scale for depression. J Neurol Neurosurg Psychiatry 1960, 23: 56-61.

32. Flemming, TR. Addressing Missing Data in Clinical Trials. Ann Intern Med. 2011;154:113-117.