This is the peer reviewed version of the following article: ‘Ratitch B, Bell J, Mallinckrodt C, Bartlett JW, Goel N, Molenberghs G, O’Kelly M, Singh P, Lipkovich I. Choosing Estimands in Clinical Trials: Putting the ICH E9(R1) Into Practice. Therapeutic Innovation & Regulatory Science (2019)’ which has been published in final form at https://doi.org/10.1177/2168479019838827

Choosing Estimands in Clinical Trials: Putting the ICH E9(R1) into Practice

Abstract

The National Research Council (NRC) Expert Panel Report on Prevention and Treatment of Missing Data in Clinical Trials highlighted the need for clearly defining objectives and estimands. That report sparked considerable discussion and literature on estimands and how to choose them. Importantly, consideration moved beyond missing data to include all post-randomization events that have implications for estimating quantities of interest (intercurrent events, aka ICEs). The ICH E9(R1) draft addendum builds upon that research to outline key principles in choosing estimands for clinical trials, primarily with focus on confirmatory trials. This paper provides additional insights, perspectives, details, and examples to help put ICH E9(R1) into practice. Specific areas of focus include how the perspectives of different stake-holders influence the choice of estimands; the role of randomization and the intention to treat principle; defining the causal effects of a clearly defined treatment regimen, along with the implications this has for trial design and the generalizability of conclusions; detailed discussion of strategies for handling ICEs along with their implications and assumptions; estimands for safety objectives, time-to-event endpoints, early phase and one-arm trials, and quality of life endpoints; and, realistic examples of the thought process involved in defining estimands in specific clinical contexts.

Key words: Estimands, Clinical Trials, Intercurrent Events

1. Introduction

Estimands and sensitivity analyses came to the forefront in 2010 with the publication of the NRC report on "Prevention and Treatment of Missing Data in Clinical Trials"[1]. An estimand describes the quantity to be estimated to address a specific study objective. It should include a definition of a population-level treatment effect measure that is defined for all subjects in the target population, including subjects that may experience certain post-randomization events during the trial. The report stressed that an estimand should be defined first and then analyses chosen in alignment with the estimand. Given the NRC report focused on missing data, it was not surprising that subsequent discussions about estimands continued to focus on missing data and the analytical methods for dealing with them. Even though a series of publications led to a process chart for choosing estimands[2, 3, 4, 5], in practice the choice of estimands, and the consequent inferences and interpretations, were still driven by habitual choices of analytical methods from which the choice of estimand could only be implicitly inferred.

The recent ICH E9(R1) draft addendum on "Estimands and Sensitivity Analysis in Clinical Trials"[6] expands the discussion beyond missing data and tackles related complex and nuanced issues that have evolved since the original ICH E9[7]. The addendum provides a language and framework for implementing the many useful ideas that have emerged since the original ICH E9 guidance, with the aim of defining and structuring a proper scientific approach where the objectives and estimands drive the trial design and analyses, rather than vice-versa. Successful implementation of the guidance is central to the principled conduct of clinical trials. The objective of this paper is to provide additional insights, perspectives, details, and examples to help put ICH E9(R1) into practice.

Specific areas of focus in the current paper include:

- How the differing decisions made by various stakeholders, at different stages of development, can lead to focus on differing objectives and estimands.
- The role of randomization and the intention-to-treat (ITT) principle.
- Means of, and issues in, defining the causal effects of a clearly defined treatment regimen, along with the implications this has for trial design and the generalizability of conclusions.
- Anticipating and defining probable intercurrent events (ICEs) as per the ICH E9(R1), that have implications for estimating quantities of interest, as a first step in choosing objectives and estimands, as opposed to adopting the automatic "missing data" perspective and the habitual choices of analytical methods.
- Detailed discussion of strategies for handling ICEs in the definition of estimands along with their implications and assumptions.
- Recommendations for the process of defining study objectives and estimands, including estimands for safety objectives, time-to-event endpoints, early phase and one-arm trials, and quality of life endpoints in the presence of many deaths.

A companion paper[8] presents practical examples of the thought process involved in defining estimands in specific clinical contexts with a variety of possible ICEs.

When discussing causal effects of treatments in this paper, we mainly refer to causality associated with an experimental treatment compared to a control tested in the same randomized multiple arm trial, although we also briefly touch upon evaluation of treatment effect in single-arm (typically early phase) trials.

In this paper, we reinforce the ICH/E9(R1) view that clinical and decision-making considerations should precede and guide the subsequent choices for appropriate estimators. Therefore, analytical aspects related to the estimators will not be discussed here and will be the focus of a separate manuscript[9].

We begin with a discussion of foundational concepts in the original ICH E9 guidance, including randomization, blinding and ITT and how they apply in today's more nuanced discussions of estimands. Subsequent sections deal with choosing objectives, dealing with ICEs, choosing estimands, and examples of choosing estimands in realistic clinical trial scenarios.

2. From past guidance to current practice

The original ICH E9 guidance[7] was issued in 1998 and it has been the foundation for statistical considerations in the design, analysis, and reporting of clinical trials. Although the guidance focused on confirmatory trials, its core principles of seeking to minimize bias are useful in every phase of clinical development. For example, blinding and randomization are standard design features that we almost take for granted in eliminating assessment and selection bias.

The ICH E9(R1) addendum[6] emphasizes the importance of quantifying treatment effects. Treatment effect is meant here as a measurement of how the outcome of an experimental treatment compares to the outcome that would have happened to the same subjects under different treatment conditions, such as, if they had not received the treatment or had received a different treatment. This means establishing the causal effects of the experimental treatment compared to control. Typically, it is not feasible to assess the same subjects under multiple treatments in exactly similar circumstances. However, randomization provides a statistical basis for establishing causal treatment effects even when each subject receives just one treatment. This is a two-step process

wherein randomization provides the causal link between subjects and the treatment to which they are assigned, and it is assumed that subjects follow the assigned treatment so that the causal relationship can be extended to the actual taking of the treatment.

However, ICEs can compromise the causal effects associated with the actual taking of the treatment and the protection randomization affords against bias, whether or not those events are related to study treatment. Examples of relevant ICEs are premature discontinuation of randomized treatment or need for rescue treatment due to lack of efficacy or toxicity of the initially randomized intervention. Such events may or may not be due to randomized treatment but will in general affect future outcomes: They may remove the effect of originally-randomized treatment and/or introduce additional treatments, incorporating their separate pharmacological effects. Subsequent outcomes will therefore be dependent upon these treatments (or lack thereof), as well as the originally randomized ones. Although the causal link between the randomized arm and the outcome may still exist, the realized treatment sequence in the randomized arm no longer consists only of receipt of the randomized treatment.

ICEs therefore need to be considered in the description of a treatment effect because they are inherently tied to causality and the interpretation of results. Depending on clinical perspective, they may represent part of the treatment, mediators, confounders or outcomes in their own right. Relying solely on randomization and ignoring the ICEs may lead to estimating the causal effect of being randomized to a treatment and not necessarily the causal effect of receiving the treatment. The former may be of relevance for some decision-makers, while the latter is arguably of ultimate practical relevance for determining the therapeutic benefits.

The ICH E9 guidance also laid out the fundamental tenants of the ITT principle, describing it as "the principle that asserts that the effect of a treatment policy can be best assessed by evaluations

based on the intention to treat a subject (i.e. the planned treatment regimen) rather than the actual treatment given". The ITT principle has two components: what subjects to include in the analysis and what data on each subject to include. The original focus of the ITT principle was primarily on what subjects to include: The ITT principle recognized the need to avoid excluding subjects who did not adhere to the planned treatment regimen to preserve the causal link to randomization and to reduce bias that could result from using a selected subset of data. As such, ITT is a foundation for defining analysis sets, with primary inferences typically driven by analyses based on all randomized subjects.

Regarding the data on each subject to include, ICH E9 mentions the importance of including observed outcomes following protocol violations or withdrawal of treatment, "consistent with the intention-to-treat principle", but does not discuss its implications on the interpretation of the trial results. With the more nuanced discussion of estimands today that was not in place when the original E9 was implemented, this is an important area to consider[10].

One way of including all randomized subjects and their observed outcomes would be to ignore all post-randomization changes of treatment and to include all their available outcomes in the analysis regardless of treatment adherence. This approach minimizes the amount of missing data. However, it breaks the link between randomization and the treatment taken and thus may lead to difficulty in attributing causal effects to the interventions of interest. Simply reducing the amount of missing data is not a valid justification. Adopting this view would be justifying an estimand based on the ease of analysis, and not on the objectives of causal inference[11, 12]. Therefore, an important aspect of the R1 addendum, and this paper, is to define the treatment of interest (both experimental and control) – the target of inference.

This issue highlights the necessity for updated guidance and the benefits from the recently proposed study development framework[5] that begins with clearly defining objectives, which lead to clearly defined estimands, in turn informing study design and analyses that are consistent with the objectives and estimands. In Section 5.1 and Figure 1 we propose a process by which efficient choice and definition of estimands may be achieved.

3. Choosing trial objectives

3.1 Considerations for differing stakeholders

In designing clinical trials, it is important to consider the decisions to be made by various stakeholders based on the evidence generated from the trial. These decisions should drive the objectives of the trial and they depend, in part, on the stage of clinical development, therapeutic area, available treatment options, etc. Phase I/II trials usually provide a proof of concept or identify doses for subsequent studies; that is, they primarily inform sponsor decisions. Phase III (confirmatory) studies typically serve diverse stakeholders and may need to address diverse objectives[3,11]. This multifaceted nature of clinical trials is another important consideration in choosing objectives and estimands[3, 5, 11, 12]. Often, multiple estimands will be needed to address the multiple objectives needed to inform the differing decisions of the various stakeholders.

For example, regulators decide if treatments should be granted a marketing authorization and in general they do so by considering the risk-benefit to patients' health of taking the drug, with reference to available alternatives. In contrast, payers decide if and where a new drug belongs on a formulary list, whether the treatment should be publicly funded and/or the appropriate reimbursement price[11]. Payer decisions typically focus on the price and benefit of the treatment relative to existing treatments. Here, the costs and net benefits are usually based on the act of

prescribing the treatment rather than taking it. Hence, regulators and payers may have different perspectives, even though they each make population-level decisions. In contrast, prescribers make many patient-level decisions. They must also inform patients and / or care-givers of what to expect from the prescribed medication. Patients/care-givers also make patient-level decisions[11]. Both are generally interested in the effects of taking a treatment as prescribed.

In general, those making decisions about groups of patients are interested in objectives and treatment effects based on the group of patients affected by the decision – even if these effects are an average estimated from heterogeneous groups of subjects. For example, averaging the treatment effects from subjects who adhere and those who do not adhere may be more useful for payers and regulators who make decisions on a population level. However, those making decisions about individual patients may not be interested in these averages because each patient has specific comorbidities, concomitant therapies, and adherence (all of which may change over time); a treatment effect estimated from the mix of adherent and non-adherent patients does not inform them of what to expect in individual patients[11]. In addition, perception of risk and / or benefit may also differ at the individual and population levels. An individual may hear 'risk of cancer' and be deterred from using a treatment, even if that risk is low and does not preclude regulatory approval.

Even for a single stakeholder in a single trial, interest may lie in more than one objective, and therefore require multiple estimands. For example, different estimands may be needed to evaluate the treatment effect in the context of other available treatment options; different estimands may be needed for safety and efficacy. Nevertheless, it is generally expedient, especially in the regulatory setting, to designate one objective and its corresponding estimand as primary, with other estimands being secondary or supportive. Regardless of the role of the estimand, defining it clearly helps ensure that the relevant evidence is obtained, properly analyzed, and interpreted.

3.2. Defining the treatment regimen under evaluation

Clearly defining the treatment (target of inference) is essential in providing the information needed by various stakeholders. For example, in proof-of-concept studies or more generally whenever assessing the effects of taking a drug, the treatment regimen of interest is typically the initially randomized treatment taken through a certain time point without any modifications or addition of other treatments. Alternatively, if there is need to incorporate the effects of efficacy and tolerability, or to assess jointly the effects of multiple therapies, then the targeted regimen may be multi-period and/or multi-treatment . In these instances, descriptions should include doses or dose ranges for the initially randomized treatments and background therapies, allowed and excluded concomitant medications, along with allowable rescue medications and their doses.

An example of a multi-period regimen is when subjects initiate an experimental drug or placebo at randomization but may discontinue treatment prior to the intended endpoint due to emergence of some contraindication (adverse reactions or lack of a response). The absence of alternative treatment options and/or the expected future course of illness may result in a plan not to treat such subjects for the remainder of the study. In this case, the treatment regimen potentially consists of the randomized treatment for a specified duration or until contraindication, with no other treatment taken after contraindication.

An example of a multi-treatment regimen is when subjects start with an initial randomized treatment but can switch to or add rescue medication if a minimal improvement is not achieved within some specified period. In this case, the multi-treatment regimen is the experimental drug plus rescue versus placebo plus rescue, with specific conditions for when rescue should be initiated and for how long it should be taken.

The strict ITT (as randomized, treatment policy) approach is one specific example of a treatment regimen. Here, the treatment regimen is the offer of the initially randomized treatment with discontinuation of that treatment and / or use of any other treatment allowed at any time without pre-specification and restrictions. It is therefore the act of randomization rather than the treatment taken that is causally linked to the outcome. This objective (target of inference) may be appealing in certain circumstances, such as for payers who want to link outcomes to financial costs. However, the target of inference is not a specific treatment regimen, nor is there guarantee that the regimens used in the trial will mirror general clinical practice where patients are neither randomized nor blinded to regimens.

Therefore, a detailed definition of the treatment regimens to be evaluated is essential because it lays out the expected course of treatment, including acceptable post-randomization flexibilities.

Some decision-makers' interest may be limited to the effects of the initially randomized treatment, while others may be interested in the entire regimen. Which treatment regimen is relevant for each objective should be assessed considering the clinical context, including natural evolution of the condition, the symptoms being treated, availability of alternative treatments, their mechanism of action and their effectiveness. The treatment regimen is the "cause" to which the outcome is attributed, with no need or intent to deconstruct the causality of the inferred effect into further pieces for the purposes of decision-making. This perspective is typical of most (randomized) clinical trials, while other approaches can be found in the causal inference literature. For example, dynamic treatment regimens and Sequential Multiple Assignment Randomized Trials (SMART)[13] may focus on inference about individual treatment components and optimization of treatment over time based on evolving outcomes.

Clarity in defining the treatment regimen provides the foundation for understanding which ICEs constitute a deviation from the regimen and how to deal with them. Clearly defining the treatment regimen is also essential for determining an adequate sample size because the assumptions about the expected treatment effect size depend on it.

4. Intercurrent events and strategies for handling them

4.1. Defining intercurrent events

ICEs are events that occur after randomization that alter the course of the randomized treatment during the intended study treatment period and/or may render subsequent endpoint measurements irrelevant. Therefore, ICEs are inherently connected to the definition of the treatment regimen. Some ICEs may be part of the treatment regimen of interest (target of inference) and therefore do not affect inference about it. Nevertheless, such ICEs should be described in the estimand definition. Other ICEs are a deviation from the regimen and must be dealt with in some manner, depending on the estimand of interest. As such, appropriate analyses are driven by the evaluated treatment regimen and the estimand of interest.

ICH E9(R1) provides several examples of ICEs; use of an alternative treatment (e.g. a rescue medication, a medication prohibited by the protocol or a subsequent line of therapy), discontinuation of treatment, treatment switching, or terminal events such as death[6]. While many ICEs are common across trials, some are specific to certain types of trials, such as incorrect food intake or vomiting in pharmacokinetic-based trials. The addendum also distinguishes between ICEs representing non-adherence to treatment and ICEs representing a change in subject's state. For example, if a subject dies in a clinical trial where death is not the primary endpoint, the subject's state changes so that the originally planned assessments will be incomplete, but further

observation is neither possible nor relevant. In contrast, non-adherence to the defined treatment regimen is an ICE which does not, in principle, preclude further observations. We refer to these circumstances as deviations from the planned treatment regimen. Missingness of data that were planned to be collected is not an ICE in itself but may be a consequence of an ICE that renders further collection of data either impossible or irrelevant for inference about the treatment regime of interest.

Both changes in state (e.g., death) and deviations from the planned regimens are ICEs that must be anticipated. We first consider deviation from the randomized treatment, and then discuss changes of subject state. In discussing the former, we assume that it is theoretically possible to observe outcomes after the event on the same schedule as for subjects who adhere to the randomized treatment and consider how the ICE affects interpretation of these outcomes.

ICH E9(R1) directs trial designers to proactively identify strategies to deal with all foreseeable ICEs. It is also useful to have guiding principles to use for unforeseen ICEs. A blinded review of all actual ICEs can be planned prior to unblinding at the end of the study. Precise definitions of the treatment regimen are needed to distinguish ICEs that are and are not deviations from the planned treatment.

To ground the discussion, consider the example subject profiles in Table 1. Assume that the trial has three post-randomization visits. Subjects are to take randomized treatment through visit 3 - the primary endpoint. The profiles in Table 1 describe treatment courses where various events may cause deviations from the initial randomized treatment. The second column depicts treatments at each visit: X = randomized treatment; O = no treatment (with the randomized treatment, experimental or placebo, discontinued and no other treatment started); ‡ = randomized treatment with an addition of concomitant rescue; and + = rescue treatment without the randomized

treatment. Therefore, profile 1 indicates adherence to the initial randomized treatment, with all others containing an ICE that is a deviation from the initial treatment.

[Table 1 Example subject profiles depicting various intercurrent events]

Table 1. Example subject profiles depicting various intercurrent events*.

	Visits			
	V1	**V2**	**V3**	**Treatment received**
1	X	X	X	Randomized treatment alone through the end of study (ideal adherence)
2	X	O	O	Randomized treatment at V1; no treatment at V2, V3
3	X	‡	‡	Randomized treatment at V1; randomized treatment + rescue at V2, V3
4	X	+	+	Randomized treatment at V1; rescue alone at V2, V3
5	X	‡	+	Randomized treatment at V1; randomized treatment + rescue at V2; rescue at V3
6	X	‡	O	Randomized treatment at V1; randomized treatment + rescue at V2; no treatment at V3
7	X	+	X	Randomized treatment temporarily interrupted for rescue at V2

* X = randomized treatment; O = no treatment (with the randomized treatment, experimental or placebo, discontinued and no other treatment started); ‡ = randomized treatment with an addition of concomitant rescue; and + = rescue treatment without the randomized treatment.

Profile #1 is a fully adherent subject with no deviations from the randomized treatment. All other profiles depict deviations. Profile #2 is a subject who receives randomized treatment for the first visit and no treatment for the last two visits. This profile is applicable only if it is ethical to leave subjects untreated for part of the study if they decide to discontinue the randomized treatment prematurely, or if no alternative treatments exist. Profile #3 adds concomitant rescue treatment to the randomized treatment. In Profile #4 subjects discontinue randomized treatment and switch to the rescue treatment. Profiles 5, 6, and 7 are mixtures of the previous scenarios. In profile #7 randomized treatment is temporarily interrupted and substituted by rescue, then resumed.

Randomization determines the initial treatment (X). The efficacy and toxicity (real or perceived) of X drives influences adherence to X and subsequent changes in therapy. Deviations from the initial randomized treatment lie on the causal pathway between the randomized treatment and outcome at visit 3. In all profiles but #1, the observed outcome at visit 3 is influenced by the effect of the initially randomized treatment and the effect of some other therapy (or lack thereof). Some or all of these scenarios (and others) may occur in almost every clinical trial. Therefore, it is important to consider early in trial planning the possible ICEs.

Once the treatment regimen of interest is determined, possible ICEs can be categorized as either part of the treatment regimen or a break from it. For example, if the treatment regimen is defined as the randomized treatment plus background therapy and possible adjustments to it if specific criteria are met, then adjustments in background therapy are part of the treatment regimen and outcomes observed after such adjustments are not confounded for inference about the target treatment regimen. It may be necessary to consider specific reasons for these adjustments, e.g., adverse events, lack of efficacy, or inability to supply/afford the medication.

For the treatment profiles that are a break from the treatment regimen of interest, outcomes after such ICEs are no longer outcomes that can be causally attributed solely to that regimen. ICH E9(R1) identifies four strategies for dealing with ICEs that are a break from the planned treatment. The choice of which strategy or strategies to use depends on the clinical question of interest. Premature discontinuation or changes to the initially-randomized treatment due to lack of efficacy or toxicity and initiation of rescue can be based on criteria pre-defined in the protocol. They can also occur spontaneously based on subject and investigator decisions. A strategy for handling them needs to be specified for both cases.

We discuss the strategies outlined in the ICH E9(R1) in the next section.

4.2. Brief overview of strategies for handling intercurrent events

ICH E9(R1) outlines five strategies to deal with ICEs:

Treatment policy strategy: The ICE marks a change in treatment course but is taken to be part of the treatment regimen of interest. The treatment effect targeted by the estimand is a combined effect of the initial randomized treatment and treatment modified as a result of the ICE.

Composite strategy: The ICE itself, possibly with outcome(s) before it, provides all necessary information about the effect of treatment. Data after the ICE provide no additional information. ICEs that can be accounted for by a composite strategy are usually important clinical outcomes that are considered to lead to an immediate conclusion about success of treatment.

While on treatment strategy: Outcomes up to the time of occurrence of ICE provide all necessary information about the effect of treatment. The actual duration of treatment in this case is not important for determining treatment benefit.

Hypothetical strategy: The ICE is a confounding factor for inference about the treatment regimen of interest. The objective is to estimate a treatment effect under a hypothetical scenario where confounding is removed, e.g., what would have happened if the ICE had not occurred; or, what would have happened under the treatment regimen of interest, which differs from the actual treatment after the ICE.

Principal stratification: The ICE is a confounding factor for inference about the treatment regimen of interest or renders future observations impossible or irrelevant (e.g., in case of death). The objective is to estimate a treatment effect in a subject population ("stratum") whose status with respect to the ICE would be identical, irrespective of treatment group; generally, the "stratum" will be that in which the ICE would not occur under randomization to either treatment.

As a brief introduction, consider profile #4 from Table 1 for a subject who discontinued the randomized treatment after one visit and received a rescue therapy for the remaining two visits in the study. Table 2 shows in bold in the second column the value(s) that may be used when applying each of the five strategies in this case.

[Table 2. Example of five strategies applied to a specific subject profile with an ICE.]

Table 2. Example of five strategies applied to a specific subject profile with an ICE.

Actual treatment profile V1 V2 V3	Target treatment regimen and relevant outcome V1 V2 V3	Strategy
X + +	X + **+**	Treatment policy
	X F **F**	Composite
	X ?O **?O**	Hypothetical
	[?] X X **X**	Principal stratification
	X – –	While on treatment
X = randomized treatment; + = rescue; F = treatment failure; O = no treatment; – = irrelevant outcome; ? = hypothetical outcome or restricted population (as in principal stratification).		

When the treatment policy strategy is chosen for this example, the observed outcome at visit 3 is measured after use of rescue therapy. In this case, the estimand compares the experimental treatment plus rescue to the control treatment plus rescue. The target of inference is the treatment regimen that includes rescue and the initially randomized treatments. When the composite strategy is chosen, the ICE itself contributes to the V3 endpoint of success/failure, and subjects requiring rescue are considered to have failed with the initially randomized treatment. With the hypothetical strategy, the hypothetical outcome of interest at visit 3 may be the outcome if no treatment (including no rescue) was provided after visit 1. Under this strategy the combined effect of

randomized treatment and rescue is not the target of inference: rescue may need to be provided in a trial for ethical reasons, but in this approach the effect had rescue not been provided is of primary interest. With the principal stratification strategy, subjects with this profile (likely to require rescue) should ideally be identified prior to study entry and not be randomized to the study. If they were randomized, then the inference would be restricted to the stratum that would not require rescue on either treatment. The question mark in column 2 of the "Principal stratification" row in Table 2 indicates that it's not immediately clear who such subjects are, and this would need to be determined either through modeling or special trial designs. With the while-on-treatment strategy, outcomes prior to the initiation of rescue are assessed (e.g. their rate or linear slope if more than one observation is available). This strategy ignores duration of treatment, the need for and timing of rescue, and focuses on the effect of randomized treatment on outcome measured up to the time of any treatment modifications.

The following subsections provide detailed discussion of each strategy, their implications and when they may or may not be appropriate. As all strategies for dealing with ICEs require assumptions, we also discuss these, while drawing distinctions from those associated with their estimation. Historically, emphasis was placed on ensuring that assumptions were conservative - not biased in favor of an experimental treatment. Although conservatism is important, especially in regulatory settings, assumptions must also be biologically plausible for the trial conclusions to be meaningful and useful to decision-makers. We emphasise that multiple strategies can be combined for the same estimand; for example, with a composite approach used for some ICEs and a hypothetical approach used for others.

Estimators for all ICE strategies must include assumptions about missing data, and there is frequent confusion between estimand definition and missing data handling (method of estimation). We

continue to assume that outcomes can in principle be measured after ICEs (except in case of death), and the focus is on whether such measurements would be useful in estimating of the treatment effect of interest. In Section 6, we discuss how some of the same strategies can be used where it is not possible to observe the outcome.

4.3. Treatment policy strategy

4.3.1 General Considerations

ICEs that are handled with the treatment policy strategy are not confounding because they are part of the treatment regimen of interest. The observed values for the variable of interest at the planned time point are used regardless of whether the ICE occurred. The treatment policy approach results, in general, in comparisons of multi-treatment regimens rather than comparisons of regimens consisting of receipt of only the initially randomized treatments. Treatment policy may be used for all ICEs, or more specifically for particular ICEs, and these are described in the following sections.

Treatment policy estimands integrate both the occurrence of an ICE and the effects of it. This means that both the occurrence of the ICE and its effects must be of interest. In general, treatment policy estimands produce treatment effect sizes of smaller magnitude than under a hypothetical approach since the effects of ICE are typically common across arms; patients that experience an ICE tend to become similar. For instance, in oncology trials it is standard practice to move subjects onto standard of care treatment following tumor progression. Treatment policy approaches confound the effects on overall survival of the initially randomized treatments because both arms share post-progression treatment; thereby reducing the magnitude of the difference between arms.

Typically, active arm patients lose efficacy after an ICE, while placebo patients may gain efficacy following rescue. Further, control arm patients are more likely to require rescue due to lack of

efficacy, while active arm patients are more likely to discontinue due to adverse events. Because of this, less efficacious arms are typically made to look better, and more efficacious arms are typically made to look worse. The opposite can also be true, e.g., where rescue is administered because the active treatment is both poorly efficacious and toxic, potentially masking the fact that the experimental treatment was not satisfactory.

Sample size calculations should be based on the anticipated effect of the treatment regimen and may differ from that which would be seen for the initial treatment alone. With smaller effect sizes under treatment policy, sample sizes for a given power need to be enlarged to compensate. If initiation of rescue does not follow strict pre-defined rules, greater variance in the outcomes may also be anticipated, also increasing sample size. Ironically, use of rescue, an ethical necessity, when accounted for using the treatment policy approach can result in exposing more subjects to inferior or ineffective treatment because of the increased sample sizes.

The reduced difference between treatment regimens makes it easier to demonstrate non-inferiority or equivalence between arms. In such cases, it may be difficult to make decisions about the comparative merits of treatment regimens without separately considering the proportions of subjects who used rescue. We note that non-inferiority margins based on historical data would also need to account for the treatment policy approach, which may be difficult to do accurately.

4.3.2 Broad treatment policy approaches

Adopting a treatment policy strategy across all possible ICEs is what has become known as a "pure ITT approach" and has been advocated as pragmatic assessments of effectiveness[11, 14]. Although this seems simple, it masks complexities that impact study conduct, statistical analysis, sample

size, cost, and interpretation of results. This is also true, to a lesser extent, when applied to only some of the ICEs.

The treatment policy strategy can only be applied if it is possible, at least in principle, to observe the outcome as planned. This strategy is not applicable for death. Since death is always at least a possibility in a trial, the treatment policy approach is technically undefinable unless death is the endpoint. However, by regarding death as missing (i.e. essentially with a hypothetical strategy) this approach may still be relevant when few deaths are expected.

When used for all possible ICEs, treatment policy broadens the treatment regimen under evaluation because it includes whatever treatment is taken (or not taken) and there may be many of these observed in a trial. Such loosely defined comparisons are rarely meaningful because the target of inference is not defined and the causal links with the investigated treatment are weakened, only proving causality between randomization and outcome. Where the rate of ICEs is high, trials adopting a treatment policy strategy will have poor sensitivity and may be unable to show the actual effects of the investigational treatment. This is of particular concern where there is a high rate of ICEs unrelated to treatment, such as in many CNS trials, and real effects can easily be lost in noise created by discontinuation or rescue. It is also a concern for treatments that only provide symptomatic relief while it is being taken, since patient data on rescue therapy may show only the effect of rescue, while for patients who just discontinued, clinically meaningful benefits observed prior to discontinuation may be ignored if they are lost or significantly diminished before the time of assessment.

It is commonly argued that broad treatment policy is most valuable because it most closely resembles real-world practice. However, unless a trial is designed to mimic real-world conditions (e.g., phase IV or pragmatic trials), treatment regimens, populations and conditions usually differ

considerably from clinical practice. Clinical trials include randomization, blinding, often placebo, precisely scheduled follow-ups, additional tests/measures and interventions, specific rules for changing doses, inclusion/exclusion criteria etc.[11, 12, 14]. In clinical practice, blinding and placebo are never used. The possibility of randomization to placebo can influence response rates and reduce adherence rates and so the generalizability of results of the treatment policy estimand when paired with placebo control may be questionable. Even where a trial is designed to be as close to clinical practice as possible, randomization is not necessarily equivalent to offering, prescribing or paying for a drug, which might be the real questions of interest for stakeholders. This strategy therefore only really proves causality between randomization and outcome, and even then, only within the clinical trial conditions. To be useful, assumptions are therefore required that the treatment regimens followed by subjects in the trial are relevant to a clinically-meaningful assessment of the treatment and that they are followed in a manner that is similar to those in the 'real world'.

In general, broad treatment policy approaches are best evaluated in pragmatic real-world designs where adherence and rescue decisions are generalizable to clinical practice[11, 14]. Treatment policy objectives may increase in relevance at the post-approval stage, especially for health technology assessment.

4.3.3 Specific treatment policy strategies

Treatment policy may effectively target specific types of ICE. The simplest and least controversial cases are where treatment switches are protocol defined (both when they occur and to what treatment(s)) and there is a desire to investigate the treatment in the context of the specified regime.

Treatment policy is frequently applied to changes in background medication, such as changes in insulin dose in Type I diabetes mellitus. Here, the alterations may be very common and/or frequent,

with effects that are difficult to ascertain. These features alone often make it logistically and statistically impractical to adopt any other approach in all but the most important cases. In contrast, choice of approach for use of concomitant medications banned by the trial protocol may be more complicated as they ought to be much rarer and with more important effects.

Another common specific strategy is where treatment discontinuation is included in the treatment regimen, e.g. including scenario 2 in Table 1. The treatment regimen would be the initially randomized treatment with possibly imperfect compliance, but no other therapeutic interventions. Starting from a broader treatment policy perspective, this is either used to exclude the pharmacological effects of alternative treatments, or arises naturally when it is ethical to ban rescue treatment, or when no alternative therapies exist. This strategy can be beneficial in several situations, notably where rescue therapy is ethically required, or new effective treatments are likely to be approved in the near future but are not yet required to be used as the comparator.

This approach should not be used to incorporate the effects of adherence into efficacy assessment: In trials where discontinuations due to lack of efficacy may occur, it is common for a successful treatment to improve both efficacy and adherence. Despite this, unless discontinuation in the active arm is reduced to zero, the effect size will typically still be smaller and less significant than if discontinuations were handled by a hypothetical strategy "if subjects continued with the initially randomized treatment". In this case, combining two pieces of evidence, each favorable to the experimental treatment, i.e., better efficacy and adherence compared to control, produces a less favorable combined result than assessing them separately.

Treatment policy may also be appropriate for handling the ICE of rescue medication when rescue has minimal impact on the endpoint being investigated. For example, while receiving a therapy to prevent asthma flares, the primary endpoint, subjects may need occasional inhaler use, which

would not influence the number of flares. It can also be useful when there is a synergetic effect of the initial treatment and rescue and rescue alone does not provide satisfactory relief. An example is treatment of migraines where patients use multi-step treatment strategies, starting with one medication and adding another, which may be needed for some portion of a patient's migraine attacks.

In placebo-controlled trials subjects may cross-over from placebo to the experimental treatment or initiate an alternative treatment approved for the indication. In general, these changes of treatment are for ethical reasons, and may not necessarily reflect the clinical question of interest. If the rescue and/or treatment switching is of interest then the treatment regimen under evaluation is a multi-therapy, dynamic regimen that evolves based on observed outcomes along with subject and clinician preferences. If it is not of interest then treatment policy is inappropriate for this ICE.

4.4. Composite strategy

With the composite strategy, whether the ICE occurred or not is combined with other measures of subject outcome to form a single composite variable. This strategy was mentioned in the ICH E10, section 2.1.5.2.2, regarding trials with rescue[15]: "In such cases, the need to change treatment becomes a study endpoint." Focus in that case is on the effect of the defined treatment regimen until the point when an ICE breaks the treatment regimen.

A common example of this strategy is where the study endpoint is defined as a responder/non-responder variable based on clinical measures or patient-reported outcomes, e.g., when a certain threshold for improvement from baseline is required. Subjects with an ICE, such as premature discontinuation of study treatment or need for rescue, are classified as non-responders. The observed outcomes determine response status for subjects who did not report an ICE.

Composite strategies are appropriate when ICEs are clearly bad outcomes, such as death or serious adverse events. However, composite strategies are more problematic for ICEs that are confounding events for the trial. For example, if early termination of study drug resulted from a serious adverse event, classifying subjects as non-responders (treatment failure) makes sense. However, if the subject discontinued because they were overly burdened by trial participation, calling them a treatment failure may not make sense. As another example, classifying a subject as a treatment failure because they needed rescue medication makes sense if the subject was not improving. However, if rescue was initiated because of subject's discomfort with the possibility of placebo treatment without adequate evidence of treatment inefficacy, it would not be appropriate.

Missing data is not an ICE, it is a consequence of ICEs. Therefore, defining all missing data as bad outcomes has no clinical justification. Composite methods must therefore be based upon ICEs rather than missingness, and hence still require a strategy to handle missing data. Imputing all missing data as a non-response is a possibility but requires strong assumptions. Composite strategies should typically also be combined with other strategies to deal with ICEs that do not represent failure.

Combining the occurrence of ICEs with numerical outcomes is less straightforward. Sometimes an unfavorable value can be ascribed as a surrogate for subjects with an ICE so that they can be included in the population-level summary of the numerical variable. The choice of this surrogate value and the corresponding summary measure is important so that the distribution of the composite variable is not unduly skewed, biased or its variance inflated. For example, in an endpoint that typically ranges from zero to ten, with low values representing bad outcomes, imputing failure as 0 may not cause statistical issues and may be clinically reasonable. In contrast, if the typical range is 5-12 and higher values are worse outcomes, imputing even a single failure

at 20 leads to inflation of variance and greatly reduced power; it may also be clinically difficult to justify that value. It may be necessary to use robust or non-parametric statistical methods to avoid such issues.

If all subjects with an ICE are assigned the same outcome, we assume that the outcome is equally bad for all such subjects and that partial effects prior to the event are irrelevant. This assumption may lead to an estimand of interest for certain treatments or indications but may not be of interest in all cases. Alternatively, it is possible to transform all measures to ranks, e.g., both the observed numeric outcomes and ICE outcomes. ICEs can be ranked based on timing, severity of adverse events, degree if insufficient response leading to initiation of rescue, etc. Ranking may often be more clinically justifiable than specifying arbitrary numerical values, and allows for more nuanced differentiation of patients with ICEs.

Composite strategies in conjunction with continuous endpoints can lead to challenging questions around the estimation of effect sizes. This is addressed further in the companion paper on estimation[9].

The composite strategy can impact sample size requirement in two ways. First, if the underlying numerical measure is dichotomized to accommodate a composite outcome or if a less sensitive summary measure is used, sensitivity is lost and larger sample sizes will be needed. Second, if composite approaches use arbitrary numeric values to represent treatment failure in case of ICEs, their choice may affect the effect size and lead to smaller or greater sample sizes depending on the rates of occurrence of ICEs. The direction and magnitude of these effects can be difficult to anticipate, and thus the power of a clinical trial with such a primary estimand may be difficult to estimate for a given sample size.

Composite approaches implicitly assume that adherence decisions in the trial are similar to those in clinical practice[11, 18]. This consideration is especially important in trials with placebo, randomization, and / or blinding because these factors are never present in clinical practice[11, 14, 18]. In placebo-controlled trials the rates of discontinuation for active treatments may be higher than when the same treatments are tested in blinded trials not including placebo. 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[11,14,18].

4.5. Hypothetical strategy

4.5.1 General Considerations

Randomized clinical trials are designed as controlled experiments to evaluate causal treatment effects under specific conditions. For various practical reasons, it is not always possible to ensure the desired conditions for all subjects. Hypothetical strategies are used to relate experimental observations to the scientific question of interest under these circumstances, via scenarios either based on some statistical model or on pre-planned sampling from observed data.

Hypothetical strategies envisage scenarios where ICEs prevent observation of the outcome under the intended conditions. They remove the effects of treatment changes that would confound estimation of the treatment effect of interest. Hypothetical strategies may translate observed trial results that are not entirely aligned with the question of interest into an answer that is aligned.

Care is required to precisely describe the hypothetical conditions reflecting the scientific question of interest in the context of the specific trial[6]. The description should contain the details of what is

hypothesized to occur and not occur. This scenario must be clearly-specified, clinically interpretable, and relevant for the results to be meaningful[6].

For example, premature discontinuation of the initial treatment may trigger two types of ICEs: no further treatment given or rescue therapy administered (see profiles #2 and #4 in Table 1). One hypothetical scenario for profile #2 is "if subjects remained adherent", whereas for profile #4 interest might be in "if subjects discontinued the randomized treatment and received no further treatment".

In general, whenever need exists to assess efficacy and adherence/safety separately, hypothetical strategies are useful. However, implementing them typically requires unverifiable assumptions. Assumptions around hypothetical scenarios may be addressed by the use of in-study data or historical data from, e.g. randomized withdrawal studies. An example for an estimand that focusses on treatment effect *had no rescue been administered* might be to use available outcomes from subjects who switch to no treatment to model outcomes for subjects who are censored when they switch to rescue. The need to justify assumptions is not unique to hypothetical strategies and occurs in any analysis with missing or irrelevant data, as discussed in Section 6.

4.5.2 Adherent Treatment Effect

The common scenario of "if subjects remained adherent" is sometimes criticized because it is counter to the fact that some subjects were not adherent and is incompatible with ethical conduct because subjects cannot be forced to adhere[11, 16]. However, the question of 'what does taking the drug do to patients?' is extremely important[11, 17, 18]. A hypothetical approach may be needed because we cannot conduct the experiment that would fully answer that question. We therefore

abstract from the observed experiment to the question of interest via appropriate analytical methods.

The effect of taking treatment as directed is of interest in many situations, including in earlier phase or mechanistic trials where interest is in a pharmacological treatment effect that *could* be achieved with ideal adherence[11, 18]. This does not preclude separate assessments of adherence and tolerability, nor considering ways for improving adherence. Another example is in non-inferiority or equivalence trials that must be designed to conform with the prescribing guidelines of the active control. It is therefore of interest to evaluate the efficacy of both the control and experimental treatment "if taken as directed". Potential differences in tolerability of the two treatments can be assessed separately. For later phase trials assessing biomarkers rather than outcomes, such as HbA1c in diabetes or the FEV1/FVC measures of lung-function in various respiratory trials, the effect of taking treatment as directed on the biomarker is of interest, particularly since further real-world outcome evidence is usually required for, or after, regulatory approval.

For those who make decisions about individual patients, knowing what happens if patients adhere may be more relevant than the effects in a mix of adherent and non-adherent patients[11, 18]. In contrast, for those who make decisions about groups of patients, the counterfactual nature what would happen if all patients adhered may not be relevant. This is not true in all cases, however. Where high background rates of subjects dropping out of clinical trials are expected (e.g., Alzheimer's or many other neurological conditions), this is often for reasons unrelated to treatment. Here, a hypothetical strategy is useful to remove confounding effects of discontinuations that would otherwise hinder the ability of a trial to assess the impact of a treatment. Even if poor adherence is expected (for all treatments) in real clinical practice, moving away from the 'real world' could actively enhance the scientific content of the trial and the ability

to perform a risk-benefit analysis by assessing them via separate analyses before the distinct signals are blurred in a combined result.

Even where interest is primarily in estimands based on outcomes associated with treatment actually taken, secondary interest may be in effects if taken as directed[11, 18]. For example, even in outcome trials (e.g., focusing on overall survival), secondary interest may be in the effect of taking the treatment, ignoring the effects of other medications. This is highlighted by the common situation in oncology trials where the currently-standard treatment policy strategy for assessing overall survival (OS) also includes the effects of changing treatment following progression, which is ethically required. Adopting a hypothetical strategy to estimate the effect on OS in subjects had treatment discontinuation or switching not been permitted allows a clearer assessment of the benefit of the experimental treatment.

4.5.3 Specific Hypothetical Scenarios

The hypothetical scenario of "if subjects discontinued the randomized treatment and received no further treatment" may be chosen to deal with ICEs of switch to rescue therapy. This strategy is useful when incorporating the effects of treatment tolerability but without including the pharmacological effects of rescue therapy. It is often unethical to withhold rescue treatment, even though its use confounds the effects of the initial medications, which are often the focus of interest.

Another application occurs where rescue medication is added to the initially randomized treatment concomitantly. An example is in placebo-controlled Type 2 diabetes mellitus trials where it is not ethical to allow prolonged insufficient blood glucose control. Assessing an effect without the concomitant rescue intake may be of interest because in clinical practice a multi-drug regimen is not preferred. In this example, the randomized treatment may already be a combination with

background metformin treatment and post-randomization rescue leads to a three-drug regimen not normally preferred in clinical practice. In this case, two hypothetical scenarios may be of interest: "if subjects continued taking the randomized treatment as planned without rescue", or alternatively, "if subjects discontinued the randomized treatment and continued taking rescue alone".

Hypothetical strategies can be limited to particular reasons for ICEs. For example, not all premature discontinuations of the initial randomized treatment occur for treatment-related reasons. In such cases, it is reasonable to consider what would happen if subjects continued with treatment, provided that the reasons for discontinuation are reported accurately. This is especially relevant in long-term studies where attrition due to study fatigue and assessment burden may increase with time.

A hypothetical strategy may therefore be required in conjunction with either the treatment policy or composite strategies. For example, when treatment policy is chosen to deal with ICEs, subjects are supposed to remain in the study and be assessed after ICEs occur. However, some subjects may withdraw from the study and the hypothetical strategy may be applied to such cases. The hypothetical scenario here is "if the subject remained in the study after experiencing the ICE" assuming the subject would undergo a treatment (or no treatment) similar to other subjects with the ICE, who remained in the study.

One scenario where the hypothetical strategy is arguably less meaningful is when the ICE is death. Although it is often numerically possible, estimating quantities for subjects who are dead is typically irrelevant. For trials with few expected deaths, unrelated to the study disease, the interpretation of applying such an approach remains acceptable[16]. This is also typically the approach taken by treatment policy so as to maintain estimability. An alternative for assessing the

effect of taking a drug while still handling interpretational issues around death is the hybrid hypothetical-composite approach. Here, most ICEs could be dealt with hypothetically but death considered an outcome. (See Section 5.2.4).

4.6. Principal stratification strategy

When interested in estimating the effect of taking only the randomized treatment, an alternative to hypothetical strategies is to adjust for ICEs. However, it is not appropriate to adjust for post-randomization variables in the same way as adjusting for baseline characteristics because of selection bias. Principal stratification is a framework in which subjects are classified into principal strata in a way that is not affected by treatment assignment, and consequently the strata can be used similarly to pre-treatment stratification variables[19].

With two randomized treatments, four principal strata can be defined with respect to a specific ICE; subjects who would experience the ICE on both treatments (A), subjects who would not experience the ICE on either treatment (B), and subjects who would experience the ICE on one treatment but not the other (C, D). Comparing outcomes between arms *within* each stratum (where occurrence of the ICE under assignment to each treatment is the same for all subjects) yields a causal treatment effect.

ICH E9(R1) suggests that the target population of interest may be the principal stratum in which an ICE, such as failure to adhere to treatment, would not occur[6]. An example where this strategy may be useful is when interest is in the effect of treatment in subjects who would be able to tolerate that treatment (and the control). With respect to rescue medication, a principal stratum could be subjects who would not require rescue medication. Another example is evaluating a causal effect

of vaccine on viral load in "subjects who would be infected", i.e., would become infected regardless of randomization to placebo or vaccine.

A further application of principal stratification is when dealing with death where death is not the endpoint of interest; for example, when evaluating the effect of treatment on quality of life where a non-negligible proportion of subjects die before the time point of interest[20]. In this case, a meaningful treatment effect referred to as Survivor Average Causal Effect (SACE), which is well defined in the principal stratum of subjects who would have survived to a specific time point regardless of which treatment they were assigned to.

In a parallel-group design it is not possible to directly observe which stratum the subject belongs to because we observe what happens only on the treatment to which the subject was randomized. The principal stratum of subjects who would not discontinue from treatment regardless of which treatment they are randomized to is not the same as a subgroup of subjects who completed randomized treatment in a parallel-group trial[6]. A completers analysis from a parallel group trial compares outcome on two different populations; using the previously defined groups, it compares subjects from strata B and C versus B and D. This may represent healthier subjects who were able to complete on placebo versus less healthy subjects who needed active treatment to complete. Unbiased comparisons require treatments to be compared on the same population[6].

Membership of a principal stratum must therefore be inferred from pre-randomization covariates, as pointed out by ICH E9(R1) – usually imperfectly[6]. This would require statistical modeling, just as modeling is necessary for hypothetical strategies, although the methodologies are different. Importantly, the original randomization is not fully preserved in principal stratification because the analyzed groups are subsets of randomized subjects. The modelling attempts to maintain

causality by attempting to achieve balance between treatment groups in the probability of having/not having the ICE.

The principal stratum strategy therefore defines the population in the estimand definition. Some trial designs can facilitate the identification of subjects in the population; such design features include cross-over, enrichment, run-in periods, and randomized withdrawal. The relevance of principal stratification for future clinical practice also should be considered: Prescribing physicians must determine whether their patients match the profiles of the clinical trial population before they prescribe the treatment. However, where the target population was not identifiable up-front in the clinical trial, physicians are also likely to be unable to define which strata a patient belongs to. It is then not sufficient to only provide the evidence of benefit in the target population and this needs to be supported by both the probability of being within it, and the benefit (or harm) if the patient does not fall within it. The translation of causal effects from a principal stratification approach to risk-benefit, or cost-benefit, assessments of taking or providing the treatment can therefore be very difficult, both at individual or population levels.

In general, principal stratification is most useful for either mechanistic goals, or in clinical cases where strata misidentification has very low cost. For instance, if it is possible to identify shortly after treatment commences that it is ineffective, then treatment may rapidly be changed. The principal stratification strategy may also be valuable as a supportive estimand if it is desirable to distinguish treatment differences associated with a specific type of ICEs vs other treatment differences.

Another situation where it is particularly important to consider the clinical relevance of principal stratification is in dealing with the ICE of early discontinuation in placebo-controlled trials. The principal stratum of interest is subjects that would adhere to both placebo and the experimental

drug. The stratum of subjects who would adhere to placebo may be a less severely ill subset and not ideal candidates for treatment with the experimental drug. Similarly, when comparing an experimental drug whose aim is improved tolerability to a standard of care, it will not be useful to focus on the stratum of subjects that would adhere to both drugs.

4.7. While-on-treatment strategy

The "while on treatment" strategy may be considered when the response to treatment before or at the time of an ICE is relevant, whereas the duration of treatment or response at a specific time point is not of particular interest. An example is palliative pain management for terminally ill cancer subjects. The treatment under investigation is not intended to prolong life, but rather improve symptoms while the subject is alive. In this case death is the ICE for which this strategy would be appropriate. However, other types of ICEs related to efficacy or tolerability of the study treatment while the subject is alive may require a different strategy.

The most common applications of the "while on treatment" strategy are where the summary measure chosen is a quantity that is independent of time. This approach can then reflect the response of subjects until the occurrence of relevant ICEs without any need to consider responses after them, in a way that handles subjects with different lengths of follow-up equivalently. Two common examples of this are rate of change in continuous endpoints where an approximately linear slope can be assumed (e.g. with the FVC or FEV1 measures of lung function) or assessment of recurrent events where a constant rate over time is assumed (e.g. with exacerbations of COPD). For these specific cases the on-treatment hypothetical and while-on-treatment estimands are also essentially identical since the required 'linearity' assumption ensures that the quantities they measure are the same.

Safety analyses are often performed on a while-on-treatment basis where the definition of on-treatment is extended to include a residual effect period. Rates of adverse events per unit exposure are typically reported, which implicitly assumes a constant hazard over time, and hence are appropriate time-independent quantities. The suitability of this strategy rests on the irrelevance of treatment duration and it essentially ignores the timing of the ICE.

When the treatment effect must be measured at a specific time for all subjects, the "while on treatment" strategy is not possible. As such, it is not equivalent to the statistical method where the last observed on-treatment value was substituted for the unobserved value at endpoint. The while on treatment approach produces the same estimate, but the interpretation reflects that the measurements were taken at different time points and hence would likely not be clinically-meaningful if interest is in treatment effect at a specific time point.

4.8. Assumptions behind the strategies for intercurrent events

All strategies for dealing with ICEs require assumptions. Some of these assumptions are directly related to the strategy itself while others arise as a consequence of the approach used to estimate the estimand. Focus here is on assumptions regarding the strategies. Estimators for all the ICE strategies must include assumptions about missing data, either as a consequence of truly unobserved outcomes or because the available data is not relevant, with brief mention of assumptions associated with the estimators. More details on assumptions for estimators can be found in our companion paper[9].

Historically, emphasis was placed on ensuring that assumptions were conservative - not biased in favor of an experimental treatment. Although conservatism is important, especially in regulatory

settings, assumptions must also be biologically plausible for the trial conclusions to be meaningful and useful to decision-makers.

4.9. Risk Benefit Implications

When evaluating the risk-benefit profile of a treatment, it is important to keep in mind the estimand and its strategies for ICEs. If use of the treatment policy or composite strategy results in outcome measures that combine efficacy with adherence and/or safety, and if this combined measure is then additionally evaluated in a risk-benefit assessment against a separate measure of safety and/or adherence, this process may result in either double-counting the risks or incoherent conclusions (see Section 4.3). In this case, a separate evaluation of safety/adherence should be used as the means to elaborate on the extent to which the composite measure is influenced by these aspects, rather than a risk counterpoint to the composite effectiveness assessment.

To avoid this double counting, the hypothetical strategy of the effect of taking treatment could be used for efficacy and then separately one could consider the proportion of subjects who discontinue the regimen.

Principle stratum also presents issues with risk-benefit (and cost-benefit) assessments since its causal effect is typically derived from a subpopulation who would not discontinue treatment, but this will typically not be knowable before the start of treatment. A full risk-benefit assessment however has to consider all patients that would be treated in clinical practice, including those who start treatment but later discontinue.

5. Choosing estimands

5.1. Defining estimands

The NRC expert panel report on missing data stressed the importance of clearly defining the primary estimand to define the target of estimation needed to address the scientific question of interest posed by the trial objective[1]. Clearly defined estimands lay a foundation for specifying aspects of trial design, conduct, and analysis needed to yield results that inform the stakeholder's decisions. General, high-level definitions of primary objectives such as "to evaluate the efficacy and safety of intervention X in subjects with condition Y" are not adequate[5].

A study development process that includes effectively defining estimands evolved in a series of publications built upon the NRC expert panel report[2, 3, 4, 5]. This process begins with considering the objectives of the trial, which entails considering the decisions made by the various stakeholders from the trial results[11]. Subsequent steps include defining the primary estimand, followed by determining design, analysis, and sensitivity analyses. The primary estimand should balance succinctness with providing sufficient detail for all relevant aspects of measuring the treatment effect. The language should be understandable by clinicians and statisticians[1].

ICH E9(R1) lists the following aspects that together describe the estimand[6]:

A. the population, that is, the patients targeted by the scientific question;

B. the variable (or endpoint), to be obtained for each subject, that is required to address the scientific question;

C. how to account for intercurrent events to reflect the scientific question of interest.

D. the population-level summary for the variable which provides a basis for a comparison between treatment conditions

The elements in the estimand definition are interrelated and need to be coherent. Specification of one element may influence the choice of the other[5, 11]. For example, the choice of a variable (endpoint) influences choice of the population-level summary that is appropriate for variable (e.g., means versus proportions). The strategies chosen to account for ICEs need to align with the variable and population. Descriptions of these estimand elements must align with the specific treatment regimen under evaluation.

Figure 1 is a more detailed version of the PSI/EFSPI[5] study development framework. The greater detail allows adaption of the framework to the specific issue of ICEs as outlined in ICH E9(R1)[6]. The following text provides additional details on the specifics of steps 1 and 2. Our companion paper covers Steps 3 and 4[9]. Previous work has stressed the need for the study development process to be iterative to account for the interrelatedness of the items[3, 5]. Although iteration may be necessary, the goal is to be as complete as possible during each step, thereby minimizing or avoiding the need for it.

- **1a**: Identify who will use the trial results and what decisions they will make from those results. This is an essential first step because the estimand must align with the decision-maker(s) needs. Any one trial may need to address the diverse needs of multiple stakeholders, leading to the need for multiple objectives and estimands.
- **1b**: Consider the broad question(s) of interest to the decision-maker. This will typically include factors such as the initially randomized treatments being compared, patient population, endpoint and time scale.

- **2a**. List all the ICEs that are plausible to occur in the trial, noting their likelihood of occurrence. Doing this early in the process avoids overlooking aspects of the intervention effect that could cause confounding or bias in analysis.
- **2b**: Define in detail the treatment regimen under investigation. The definition should include whether interest is in the effects of the regimen if taken as directed or as actually taken. It is therefore also necessary to specify whether each ICE is part of the regimen under investigation or is a departure from the intended regimen.
- **2c**: Based upon the treatment regimen defined in 2b, specify the strategy to handle each type of ICE. The ICEs that are part of the regimen are handled using the treatment policy strategy. For those ICEs that are deviations from the treatment regimen, use the clinical question to determine which ICE are outcomes (e.g., dropout due to adverse events is considered treatment failure), thus using the composite strategy to handle them. The remaining ICEs are confounding factors that are problematic for the assessment of the outcome and treatment regimen of interest. They can be dealt with via a hypothetical or principal stratification strategy, or avoided, i.e., handled with the while-on-treatment strategy. To complete the definition of the estimand, define the population, endpoint, and summary measure considering the classifications of ICEs and chosen strategies to handle them. Revisit the clinical objective to check that the estimand aligns to it and that it fits the stakeholder requirements.

[Figure 1. Study development process chart. PLACEHOLDER]

Figure 1. Study development process chart.

5.2. Special considerations in defining estimands

5.2.1. Estimands for safety objectives

Safety outcomes are sometimes a primary objective of a trial, but even when not primary, safety is always important to evaluate. The ICH E9(R1) estimands framework may be applied to safety analyses[21]. As with efficacy outcomes, it is important to define the population, outcome, ICEs and summary measure; and, a treatment regimen should again be definable to attribute adverse effects correctly.

Safety assessments are frequently based on the initially-randomized treatment only, with analysis while on treatment estimand. Residual effect periods are typically included to allow for delayed reaction to treatment and/or attribute events to the drug while the active substance remains in the body. Different aspects of safety assessment may require different estimands. For example, absolute numbers of AEs and rates of AEs each provide different perspectives on safety.

Issues may arise when integrating efficacy and safety as part of a risk-benefit assessment if the two derive from estimands with different treatment regimens. If efficacy was assessed based on a treatment regimen including rescue while safety was based on the initially randomized treatment, then benefits and risks are not directly comparable; the benefits of rescue are included without the drawbacks.

5.2.2. Estimands for early-phase trials

The estimands framework applies to all clinical trials because it is always necessary to define what is to be estimated. However, nuances and different considerations apply to certain types of trials. Estimands for early-phase trials may differ from the confirmatory settings focused on in ICH E9(R1) in several ways, most prominently in regards to the stakeholder, and in some instances the use of single-arm trials.

Some early phase trials are designed with regulatory stakeholders in mind (e.g., oncology trials) because they may form part of the submission package. These trials may focus on estimands typical of confirmatory trials. However, most early phase trials are designed to inform later trials and to inform the sponsor about whether further investment in the drug was warranted[11]. Early phase estimands focus less on treatment policy; rather the estimand tends to address narrow definitions of the treatment regimen, to establish proof of concept and to maximize the probability of making correct decision about advancing compounds to later phases. However, collecting data on adherence, ICEs and what happens after ICEs may still be important secondary objectives for the planning of later trials.

A second key difference between early and late phase trials is that sometimes early phase trials are single arm, that is, have no control arm. This is also common in long-term extensions of late-phase trials. Estimands are often defined as a comparison between treatments within the trial, but this is not possible in this setting. Instead, estimands need to focus on single-arm estimation and/or on comparison to a pre-defined (or historical) target. Causality is harder or even impossible to establish in single arm trials, but nonetheless most of the concerns about the different methods of dealing with ICEs remain. In general, approaches based on the hypothetical and/or composite strategies will be most straightforward to compare to results from other trials as their estimates

will be based on more clearly defined scenarios and thereby better suited to like-for-like comparisons; or, if such comparisons are not possible, then it is easier to identify how conditions differ.

5.2.3. Estimands for evaluations of time to event

For time-to-event trials such as oncology, the recommendations and considerations in ICH E9(R1) apply[22]. Here, ICEs may be competing events; for instance, 'other death' in an assessment of cardiovascular death. The protocol may define withdrawal of randomized treatment, such as after tumor progression in oncology. Traditionally definitions for censoring in the analysis have involved a mixture of ICEs. The treatment regimen definition is also critical in what are often trials of long duration with multiple rescue treatment options. In oncology interest may exist in questions regarding the sequencing of treatments, or whether treatment may be stopped successfully after a certain period. When this is the case, the importance of a clearly defined treatment regimen is essential.

One of the biggest challenges is that there are potential delays between treatment and outcome, and that intervening ICEs are often a sign of lack of efficacy. Examples of this are in oncology with overall survival, or in general where short term disease-modifying treatments have long-term outcomes (e.g. surgery). Analysis based only on short-term data prior to the ICEs can be biased with respect to long-term effects of the initial treatment. Historically, the solution has been to use treatment policy approaches that include all available data to infer effects even after changes of treatment. However, this is another example where attenuation of the initially assigned treatment's effect is likely and leads to loss of the causal link between taking treatment and outcome. When focus is on the causal effect of taking treatment, a hypothetical approach addresses the question directly. Typically, the hypothetical strategy requires more complex statistical methods, e.g., rank-

preserving structural failure time models for survival data[23, 24, 25] that entail more assumptions, particularly around informative censoring. However, this trade-off may be justifiable because in these cases the simpler treatment policy approach is not estimating the effect of the treatment regimen of interest.

5.2.4. Estimands for quality of life evaluation in trials with many deaths

In trials where many deaths are anticipated, such as oncology or heart failure studies, the evaluation of quality of life (QoL) is challenging. QoL scales assess aspects of patient satisfaction and function while alive and do not have a designated value or category for death. Several potential questions of interest exist for these outcomes.

A straightforward question is QoL "while alive" without regard for the duration of survival. This approach may be of interest when assessing the QoL and survival separately or for stakeholders who place more importance on the QoL than duration of survival. An estimand for this objective would therefore use a "while-on-treatment" strategy with respect to the ICE of death and a summary measure incorporating all available measurements up to death.

Alternatively, treatment benefit at a specific time point may be of interest, e.g., one year after the start of treatment, using a combined measure of QoL and survival. In this case a composite strategy could be used for death as an ICE. However, this approach exemplifies the need for care in defining the composite endpoint. Assigning a numerical pseudo value for subjects who die (e.g., a zero value as the lowest possible QoL score) may not be the best strategy. A zero for quality of life is often regarded as not adequately representing mortality and could skew treatment differences and be difficult to interpret. Ranking or categorical approaches could instead be used based on QoL

measurements if subjects survive to the time point of interest with the worst rank/category(s) reserved for those who died (possibly also depending on time of death).

Another common composite strategy is that of Quality Adjusted Life Years (QALYs), whereby the AUC of QoL truncated by death is calculated. This reflects a combination of survival and QoL, and is commonly used as a measure of clinical utility for cost-benefit assessments by payers. Other possible related approaches include time-to-event endpoints such as time to QoL deterioration or death (whichever comes first), or QoL-weighted survival time.

To illustrate differences between estimands, consider for example, two cancer treatments. One prolongs survival but has severe side effects, while the second has minimal effect on survival but few side effects. While-alive estimands would favor the second treatment regardless of relative survival times, but the survival-QoL composite estimands would take into account the longer survival time to give a more balanced assessment; the preferred treatment would depend on the relative lengths of survival and the quantitative impact of the side effects.

Alternatively,

Another objective may be to evaluate the QoL at a specific time point of interest for subjects who survive to that time point. Here a principal stratification strategy could be used. Modeling is required to identify those subjects who would survive to the time point of interest regardless of which treatment they were assigned to. As noted above, this requires strong assumptions that survival is predictable from baseline characteristics and that all relevant covariates are measured.

5.3. Trial design and conduct considerations

Treatment policy strategies require the collection of as much data as possible for subjects even after discontinuation of randomized study treatments. Other strategies do not require this data

because it is either irrelevant (as in some hypothetical strategies) or unused (as in composite, principle strata, while-on-treatment strategies). Many strategies involve at least some element of treatment policy, and there may be multiple estimands of interest in the trial, not all of them apparent at the time of planning.

In early-phase trials focusing on proof-of-concept, it may be useful to gather information about the likely outcomes for subjects following discontinuation to inform future confirmatory trials. Even where post-ICE data are irrelevant for all reasonable estimands, such data can still provide information useful in sensitivity analyses, mainly by placing limits on what is likely to have occurred had treatment continued. For example, subjects' measurements following discontinuation could be used as a conservative estimate of their performance had they continued treatment as planned. Therefore, standard practice should be to collect post-ICE data, at least for the primary endpoint.

It is also important to collect information about the ICEs themselves[6]. For any estimand where distinction is drawn between types of ICEs, it is important that categorizations of type are accurate, pre-defined and as objective as possible. Some distinctions are based on changes in treatment and it may therefore be important to record subject medication usage and post-ICE outcomes. Other distinctions between ICEs may be made based upon the reason for discontinuation; for instance, discontinuation due to lack of efficacy may constitute a different ICE than discontinuation due to a serious adverse event. Where such distinctions are part of an estimand, these reasons should be defined a priori and objectively. Specific training on this should be provided to sites.

Uninformative reasons for discontinuation such as 'lost to follow-up' and 'withdrawal of informed consent' should be minimized through trial design and conduct. To handle the occurrence of unforeseen types of ICE, a blinded review of ICEs could occur concurrent with blinded review of

protocol violations. When categorizing ICEs, it is important both to note what the change in treatment was (e.g. change of background therapy, discontinuation of randomized treatment, use of rescue therapy), and also the root cause of why it occurred. For example, if a patient discontinues treatment, is it due to adverse events, lack of efficacy or clinically-unrelated reasons? Both the types of ICEs and the reasons for them should be reported, and where possible comparisons drawn between treatments both numerically and visually. This can be vital supplementary evidence for assessing the effect of treatment for all estimands, although its interpretation will depend upon the estimand chosen (and in particular whether these effects are already included in the estimand or not).

The treatment policy strategy also has an impact on trial conduct. The distinction between discontinuation of the initially randomized treatment and discontinuation from the study overall is fundamental to the implementation of this strategy. All subjects should continue study follow-up as originally planned even if they are no longer taking the initially randomized treatment. This needs to be specified in the Informed Consent Form and discussed with subjects in advance. If rescue medication is provided to study participants at no cost, it may motivate them to remain in the study, although this will increase trial cost. If alternative therapies are not covered for the duration of the study, the subjects will have to essentially consent to trial-related assessment burden while taking treatment that they could obtain in the same way outside of the study. Protocols may need to make provisions for a reduced assessment schedule to collect only the most essential information in these cases.

Full discussion of all the possible design considerations for the various approaches to dealing with ICEs is beyond the scope of this paper. Nevertheless, the considerations are critical. For example, when implementing the principal stratification approach, need may exist to collect a more

extensive set of demographic and prognostic baseline assessments to predict stratum membership with sufficient accuracy.

Some general principles are useful to consider when making specific design decisions for a trial. First, consider ICEs as belonging to one of two categories: avoidable and not avoidable. The NRC expert panel report on missing data[1] emphasized the importance of limiting missing data. The principle should be extended to ICEs in general. For example, if subjects need to discontinue study drug due to adverse events or lack of efficacy, so be it. However, discontinuations for other reasons and avoidable ICEs more generally should be minimized by trial design and conduct[26].

A second principle is to minimize reliance on definitions and statistical models and strive to collect data by design. For example, whenever possible use cross-over, run-in or randomized withdrawal designs rather than relying on a model-based principal stratification. Do not rely on a hypothetical strategy if the data to address the estimand of interest could be collected. Do not rely on a broad treatment policy strategy when it is possible to implement more relevant and precisely defined treatment regimens.

A third principle is to design trials to answer the actual questions of interest, not the easily answered questions. Do not rely on composite and while on treatment strategies simply because they limit missing data. It is not justifiable to change the question of interest for convenience[27].

6. Missing data

Throughout this paper, discussion of missing data has been deliberately minimized because defining an estimand is based upon ICEs rather than missing data. The estimand defines the property of the target population that is being measured, and consequently its definition is independent of missing data. Missing data is often associated with estimands because many ICEs

lead to missingness (although ICH E9(R1) encourages the collection of data following ICEs if such data are meaningful), and in estimands where post-ICE measurements are not relevant, missing data arise.

All well-defined estimands are prone to missing data since measurements may be missing even in the absence of ICEs. It is not appropriate to define an estimand whereby either missingness is an outcome, or it is substituted by an arbitrary value since missingness is not a population-level property. It may sometimes be reasonable to use non-response imputation for all missingness in a composite strategy. However, this is only a method of estimation, cannot be easily represented in an estimand.

As all estimands are at risk of missing data, all estimators (analysis methods) require untestable assumptions to handle missing data. However, by their definitions the different estimand strategies differ in the amounts of unobserved data that they may be susceptible to and in their robustness to assumptions made. The more missing data, the more sensitive estimates may be to the assumptions. In this context, hypothetical methods could be said to be dependent not on assumptions about missing data, but on assumptions that the scenario required by the estimand can be modelled from observed data due to practical/ethical reasons. Treatment policy approaches, on the other hand, tend to be sensitive to missing data, since they require measurements post-ICE which are often difficult, and in some cases impossible (e.g., post-death), to obtain. Principle strata strategies are dependent upon assumptions as, or more, severe as that of hypothetical strategies since extrapolation from a model is generally required, e.g., with a model based on one treatment group being applied to subjects from another treatment group in order to determine stratum membership. While-on-treatment and composite methods are defined so that post-ICE data is irrelevant but

missing pre-ICE data is still possible. They also require other strong assumptions such as all bad outcomes being equally bad or constancy of effect over time.

Universal agreement exists that trials should aim to maximize adherence to protocol procedures, including adherence to the protocol-assigned treatments[1]. Maximizing adherence improves robustness of results by reducing the reliance of inferences on the untestable assumptions[1]. Similarly, improving subject follow-up post-ICE improves the robustness of estimation of treatment-policy estimands. However, complete follow-up is rarely possible so for all methods sensitivity to assumptions should be checked with sensitivity analyses. The companion paper on estimation methods covers missing data handling and means of providing sensitivity analyses[9].

7. Conclusions

Although estimands and sensitivity analyses have received considerable attention in recent years, a common language for statisticians and clinicians was lacking, which limited progress. The ICH E9(R1) draft addendum provides this language and provides a framework for implementing the useful ideas that have emerged in recent years. This paper builds upon the addendum with additional discussion of the decision-making context at various stages of drug development by different stakeholders. It describes how the new process of defining estimands is expected to benefit everyone involved in evidence-generation and decision-making.

The draft addendum stresses that the backwards process of the primary statistical analysis implicitly defining an otherwise unspecified primary estimand is not acceptable because it leads to pitfalls in interpretation of results. This paper reinforces that view by focusing on the clinical and decision-making considerations that are central to the estimand definition. These considerations precede and guide the subsequent choices for appropriate estimators.

Elements of the estimand definition defined by ICH E9(R1) were elaborated upon in this paper, with detailed focus on ICEs occurring after treatment initiation that change the treatment being administered, and therefore affect the interpretation of contrasts of outcome between randomized groups. While other elements of the estimand definition (population, endpoint, and population-level summary) have always been described in study protocols (albeit not necessarily as part of the estimand), the inclusion of ICEs handling in the estimand definition is an important new requirement, which should improve clarity of study objectives.

Extensive details and insights on the strategies for handling ICEs and their implications were provided. Focus included how the strategies for ICEs relate to the intent-to-treat principle laid out in the original ICH E9. Emphasis was placed on clearly defining the treatment regimen under evaluation as a key object of decision making. Although a definition of the treatment regimen under evaluation is not explicitly listed by ICH E9(R1) as part of the estimand definition, it directly links it to the clinical question of interest. The initial (randomized) treatments and the treatment regimens that are assessed and compared through the trial are not always the same. Hence a clear definition of the treatment regimen under evaluation is essential in defining the estimand.

The inter-dependence between the elements of the estimand definition and the definition of the treatment regimen are important and may lead to an iterative process of revision and refinement to arrive at the final definition of the estimand. However, with careful planning the need for iteration can be minimized and this should foster greater clarity throughout the study development process.

We also touched upon some aspects that received less attention in the draft addendum, such as estimands for safety objectives, early-phase trials and key distinctions for defining estimands in single-arm trials and time-to-event endpoints. Examples of the thought process and considerations required to define estimands in specific clinical contexts were provided in the companion paper[8].

Although ICH E9(R1) and this paper consider estimands in the context of randomized clinical trials, the estimand concept readily applies to any study of a therapeutic intervention, whether randomized, observational, or mixed. Estimands are naturally incorporated within the framework of causal inference, e.g. using the language of potential outcomes[28]. One form of estimand is the difference in outcome had each treatment been applied to all patients from the population, even if contrary to the fact. This definition conveys the target of estimation in a very general manner and does not depend on specific mechanisms of treatment assignment. Different estimand strategies can be expressed in the language of potential outcomes with the benefit of defining estimands in a unified way across different types of clinical studies, which will be the topic of a sequel paper on causal inference.

8. References

[1] National Research Council. The prevention and Treatment of Missing Data in Clinical Trials. Panel on Handling Missing Data in Clinical Trials. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press; 2010.

[2] Mallinckrodt CH, Lin Q, Lipkovich I, Molenberghs G. A structured approach to choosing estimands and estimators in longitudinal clinical trials. ***Pharm Stat.*** 2012;11:456–461.

[3] Leuchs AK, Zinserling J, Brandt A, Wirtz D, Benda N. Choosing appropriate estimands in clinical trials. ***Ther Innov Regul Sci.*** 2015;49(4):584-592.

[4] Garrett A. Choosing Appropriate Estimands in Clinical Trials (Leuchs et al): Letter to the Editor. ***Ther Innov Regul Sci.*** 2015; 49(4):601-601.

[5] Phillips A, Abellan-Andres J, Soren A, et al. Estimands: discussion points from the PSI estimands and sensitivity expert group. ***Pharm Stat.*** 2017;16(1):6-11.

[6] ICH Harmonised Guideline E9(R1). Estimands and Sensitivity Analysis in Clinical Trials. Step 1 version dated 16 June 2017.

[7] ICH Harmonised Tripartite Guideline E9. Statistical Principles for Clinical Trials. Step 4 version dated 5 February 1998.

[8] Ratitch B, Goel N, Bell J, Mallinckrodt C, Bartlett JW, Singh P, Lipkovich I, O'Kelly M. Defining Efficacy Estimands in Clinical Trials: Examples Illustrating ICH E9(R1) Guidelines. Submitted to ***Ther Innov Regul Sci.*** 2018.

[9] Mallinckrodt CH, Bell J, Ratitch B, Liu G, O'Kelly M, Lipkovich I, Singh P, Xu L, Molengerghs G. Technical and Practical Considerations in Aligning Estimators with Estimands in Clinical Trials. Submitted to ***Ther Innov Regul Sci.*** 2018.

[10] Leuchs AK, Brandt A, Zinserling J, Benda N. Disentangling estimands and the intention-to-treat principle. ***Pharm Stat.*** 2017;16:12–19.

[11] Mallinckrodt CH, Molenberghs G, Rathmann S. Choosing estimands in clinical trials with missing data. ***Pharm Stat.*** 2017; 16:29–36.

[12] Akacha M, Bretz F, Ruberg SJ. Estimands in clinical trials – broadening the perspective. ***Stat Med.*** 2017;36(1):5-19.

[13] Chakraborty B, Murphy SA. Dynamic Treatment Regimes. ***Annu Rev Stat Appl.*** 2014;1:447-464.

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

[15] ICH Harmonised Tripartite Guideline E10. Choice of Control Group and Related Issues In Clinical Trials. Step 4 version dated 20 July 2000.

[16] Permutt T. A taxonomy of estimands for regulatory clinical trials with discontinuations. ***Stat Med.*** 2016;17:2865-2875.

[17] Mallinckrodt CH, Roger J, Chuang-Stein C, et al. Recent Developments in the Prevention and Treatment of Missing Data. ***Ther Innov Regul Sci.*** 2014;48(1):68-80.

[18] Mallinckrodt CH, Lipkovich I. A practical guide to analyzing longitudinal clinical trial data. Boca Raton, FL: Chapman and Hall/CRC; 2017.

[19] Frangakis CE, Rubin DB. Principal stratification in causal inference. ***Biometrics.*** 2002;58(1):21–29.

[20] Rubin DB. Comment on: Dawid AP. Causal inference without counterfactuals. ***J Am Stat Assoc.*** 2000;95(450):407-424.

[21] Unkel S, Amiri M, Benda N, et al. On estimands and the analysis of adverse events in the presence of varying follow-up times within the benefit assessment of therapies. ***Pharm Stat.*** 2018; https://doi.org/10.1002/pst.1915.

[22] Rufibach K. Treatment effect quantification for time-to-event endpoints–Estimands, analysis strategies, and beyond. ***Pharm Stat.*** 2018; https://doi.org/10.1002/pst.1917.

[23] Robins JM, Tsiatis AA. Correcting for non-compliance in randomized trials using rank preserving structural failure time models. ***Commun Stat-Theor M.*** 1991;20:2609-2631.

[24] White IR, Babiker AG, Walker S, Darbyshire JH. Randomisation-based methods for correcting for treatment changes: examples from the Concorde trial. ***Stat Med.*** 1999;18:2617-2634.

[25] White IR, Walker S, Babiker AG, Darbyshire JH. Impact of treatment changes on the interpretation of the Concorde trial. ***AIDS.*** 1997;11:999-1006.

[26] Hughes S, Harris J, Flack N, Cuffe RL. The statistician's role in the prevention of missing data. ***Pharm Stat.*** 2012;11:410-416.

[27] Flemming TR. Addressing Missing Data in Clinical Trials. ***Ann Intern Med.*** 2011;154:113-117.

[28] Rubin DB. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. ***J Ed Psy***. 1974;66(5):688–701.