*Statistics in Medicine*

WILEY

Statistics in Medicine

RESEARCH ARTICLE OPEN ACCESS

# A Perspective on the Appropriate Implementation of ICH E9(R1) Addendum Strategies for Handling Intercurrent Events

Thomas R. Fleming[1] | Kevin J. Carroll[2] | Janet T. Wittes[3,4] | Scott S. Emerson[1] | Mark D. Rothmann[5] | Sylva Collins[5] | Gregory Levin[5]

[1]Department of Biostatistics, University of Washington, Seattle, Washington, USA | [2]KJC Statistics, LTD, Woodford, UK | [3]Wittes LLC, Washington, DC, USA | [4]Affiliate Professor of Biostatistics, Department of Population Health and Social Science, Florida Atlantic University, Boca Raton, Florida, USA | [5]Office of Biostatistics, Office of Translational Science, Center for the Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, Maryland, USA

**Correspondence:** Thomas R. Fleming (tfleming@uw.edu)

**Received:** 2 July 2024 | **Revised:** 7 March 2025 | **Accepted:** 11 April 2025

**Funding:** This work was supported by the National Institutes of Health, R37-AI29168.

**Keywords:** causal inference | estimands | intercurrent events | randomization

## ABSTRACT

In randomized clinical trials, stopping study medication, use of rescue treatment, and other intercurrent events can complicate interpretation of results. The ICH E9(R1) Addendum on estimands stimulated important cross-disciplinary discussions on trial objectives. Unfortunately, with influence of the Addendum, many trials have proposed analyzing primary endpoints using “while on treatment”, “hypothetical”, or “principal stratum” strategies that handle intercurrent events in ways that use post-randomization outcomes to exclude information from randomized participants and don’t preserve integrity of randomization, or that don’t reliably capture the intervention’s meaningful net effects. These approaches have inherent limitations in ability to draw scientifically rigorous inference on clinically relevant causal effects important for decisions about adopting interventions. We describe advantages of trials with standard-of-care control arms targeting estimands using “treatment policy” approaches for intercurrent events, while potentially incorporating other meaningful intercurrent events, such as death, into the primary endpoint applying a composite strategy. Well-designed and -conducted trials targeting such estimands achieve scientifically rigorous causal inference through analyzes that protect the integrity of randomization. Such estimands also provide meaningful information relevant to real-world settings because they (1) are unconditional in nature i.e., they don’t condition on post-treatment circumstances that might not be many participants’ experiences; and (2) properly evaluate the experimental intervention within a regimen that includes possible ancillary care that would be clinically appropriate in real-world settings. We hope to add clarity about appropriate strategies for intercurrent events and how to improve design, conduct, and analysis of clinical trials to address questions of greatest clinical importance reliably.

## 1 | Introduction

The ICH E9(R1) Addendum on estimands [1], hereafter, the “Addendum”, grew out of the 2010 National Research Council’s report [2] concerning the impact of missing data in clinical trials. The report recommended clearly identifying the target “estimand” when designing and analyzing clinical trials to assess the appropriateness of methods for handling missing data. The

**TABLE 1** | Examples of planned primary analyzes using hypothetical, principal stratum, or while on treatment strategies for intercurrent events[a].

1. Hypothetical strategy

“The *values [for the efficacy measure] following initiation of [the rescue therapies] post randomization will be handled by the “hypothetical strategy” outlined in the ICH E9 (R1) guidance, meaning that values following the administration of these rescue therapies will be imputed so as to reflect that their occurrence likely indicates disease worsening.”*

“*Observations collected after initiation of REDACTED will not be considered in the primary analysis*”.

“REDACTED *would like to confirm [with the regulatory authority] our plans for handling intercurrent rescue events. We plan to allow subjects who receive [the rescue treatment] to remain in the study, but to carry the last observation prior to rescue forward and to censure subsequent efficacy data from the analysis.”*

2. Principal stratum strategy

*“For patients who could not tolerate [dose 1], the dosage can be reduced to [dose 2]. If you propose a further dose reduction to [dose 3], the efficacy and safety results should be analyzed separately for those patients receiving [dose 3].”*

3. While on treatment strategy

“*To increase sensitivity, efficacy and safety events will be included only if occurring while on treatment or within 30 days of treatment discontinuation.”*

*“We recommend “while on treatment” strategy as the primary analysis of recurrent events because the data after ICEs that may impact efficacy should be excluded in the primary analysis”*.

[a] Examples drawn from academic authors’ experiences, with redactions for confidentiality. These six quotations come from six separate studies.

Addendum stressed the importance of ensuring that the chosen statistical estimator indeed estimated the chosen estimand. By providing a common language to describe how handling of missing data modifies the target of scientific investigation, the Addendum can help standardize review of protocols. In the European Union, the Committee for Medicinal Products for Human Use (CHMP) issued the Addendum as draft guidance in 2017 with final adoption in 2020. The United States FDA released the draft in 2017 and finalized it in 2021.

In defining its purpose, the Addendum indicates, “*To properly inform decision making by pharmaceutical companies, regulators, patients, physicians and other stakeholders, clear descriptions of the benefits and risks of a treatment for a given medical condition should be made available*,” adding, “*without such clarity, there is a concern that the reported “treatment effect” will be misunderstood.* [*T*]*he question remains whether estimating an effect in accordance with the ITT principle always represents the treatment effect of greatest relevance to regulatory and clinical decision making. The framework outlined in this addendum gives a basis for describing different treatment effects and some points to consider for the design and analysis of trials to give estimates of these treatment effects that are reliable for decision making*.” While the framework the Addendum discusses can potentially apply to all analyzes of clinical trial data, this manuscript focuses primarily on primary and multiplicity-controlled secondary endpoints meant to provide scientifically and statistically rigorous confirmatory evidence of a treatment’s effect on a clinically important outcome. We authors, coming from consulting, data monitoring, regulatory, and peer-review settings, have seen many clinical trials, with influence of the Addendum, propose strategies that handle intercurrent events inappropriately. Such trials use post-randomization outcomes to exclude information from randomized participants, thus not preserving the integrity of randomization, or fail to capture the meaningful net effects of the intervention. In particular, many primary analyzes target estimands using “while on treatment”, “hypothetical”, or “principal stratum” strategies for such intercurrent events as treatment discontinuation or rescue medication use, (See Table 1.) Arguments supporting use of these estimands often fail to recognize explicitly the inherent limitations in their ability to infer causality, or to acknowledge the potential to mislead physicians and patients about effects of the intervention on outcomes of most importance.

This paper emphasizes principles of fundamental importance to protecting the integrity and reliability of causal inference in randomized clinical trials. Guided by these principles, we then formulate critically important insights about how to proceed when selecting estimands. We do not simply indicate what should be done, but we present our reasoning. Several examples illustrate the limitations of some estimand strategies as well as the advantages of those that respect randomization.

We discuss in Section 2 features of trials that enhance their ability to achieve scientifically rigorous causal inference and to provide meaningful information relevant to real-world settings. We focus on randomized trials with standard-of-care control arms targeting estimands that employ a “treatment policy” strategy for intercurrent events (e.g., treatment discontinuation, failure to implement protocol-specified dosing strategies, and use of rescue medication), while potentially applying a composite strategy to incorporate other meaningful intercurrent events, such as death, into the primary endpoint. Section 3 discusses how intercurrent events shape estimand strategies, and the consequences of mishandling such events. We provide additional recommendations for appropriate use of estimands in clinical trials in Section 4 and conclusions in Section 5.

## 2 | Elements of Trials Promoting Scientifically Rigorous Causal Inference on Meaningful Estimands

This section describes, and Table 2 summarizes, specific elements of design, conduct, and analysis which enable a trial to provide

**TABLE 2** | Elements of trials promoting scientifically rigorous causal inference on meaningful estimands.

1. Randomization between experimental and control arms.
2. Use of a standard-of-care control arm.
3. Use of a primary endpoint that directly measures how participants feel, function, or survive or is an appropriate replacement endpoint.
4. Follow-up of all participants for the full intended observation period or until valid withdrawal of consent, regardless of intercurrent events such as treatment discontinuation or use of rescue treatments.
5. A primary analysis based on the ITT principle in that it includes all randomized participants (or all randomized participants in a pre-specified cohort, where such pre-specification is based on baseline covariates or samples collected at the time of randomization), avoids using post-randomization outcomes to exclude information from randomized participants, and uses a treatment policy strategy for intercurrent events such as treatment discontinuation and rescue medication use, while potentially also incorporating other meaningful intercurrent events, such as death, into the primary endpoint applying a composite strategy.

*Note:* Sensitivity analyzes to investigate and quantify the impact of missing data on conclusions that include a broad spectrum of conjectures about mechanisms of missing data that are not missing at random.

scientifically rigorous and reliable causal inference on clinically meaningful estimands. More precisely, the following elements can facilitate the reliable evaluation of the effects of prescribing an experimental intervention, when compared to prescribing a reasonable version of standard-of-care, on aspects of how patients function, feel, or survive, at actual levels of adherence to treatment and in the presence of ancillary care and rescue therapies that might be expected in a real-world setting:

1. Randomization: Randomizing study participants between experimental and control arms, ideally in a double-blind manner and proximal to the time of initiation of the randomized intervention, enables reliably distinguishing the causal effects of an intervention (i.e., the changes in outcome caused by the change in intervention) from disease-related consequences in the presence of other confounding factors. ICH E9 and E9(R1) both highlight the critical role of randomization, as does a substantial body of published material [3].

2. Use of a standard-of-care control arm: Using a control arm in which participants receive a version of local standard-of-care, while the experimental intervention is part of a regimen allowing proper supportive care, addresses the best interests of participants and provides information relevant to treatment decisions that patients and prescribers would make if the intervention were approved. This is not typical practice in many symptomatic diseases with approved safe and effective therapies, where clinical trials often include placebo control arms, withholding standard-of-care therapies, providing them only as post-randomization rescue treatment. Such designs have commonly been used to evaluate drugs in trials in dermatology (e.g., psoriasis), rheumatology (e.g., rheumatoid arthritis), gastroenterology (e.g., ulcerative colitis), and pain (e.g., chronic pain). While regulations do not require the comparator to be standard-of-care in some of these settings, these designs often introduce substantial challenges to analysis, as recognized in the National Research Council Monograph [2], because problematic intercurrent events arise when many participants indeed receive rescue medication. They also often do not provide information most relevant to real-world treatment decisions. For example, in disease settings where patients and prescribers typically choose between available therapies, a comparison against no therapy in a placebo-controlled trial does not provide results to inform most directly the treatment decisions made in clinical practice. Application of standard-of-care control arms could be facilitated in many such settings, such as rheumatoid arthritis and psoriatic arthritis, by carrying out active-controlled non-inferiority trials or add-on superiority trials [4]. Also, when there is interest in understanding efficacy in a clinical setting where a certain supportive care medication is not generally provided, in some cases a trial can be conducted in selected sites or regions that have little or no expected use of that additional medication prior to the assessment of the primary endpoint.

3. Use of a primary endpoint that directly measures how participants feel, function, or survive [5] or is an appropriate replacement endpoint: Using such endpoints ensures that evidence of a treatment effect establishes or reliably predicts a direct benefit to patients in terms of how they feel, function, or survive. As discussed below, in some cases a composite strategy may be used to evaluate a composite endpoint that incorporates meaningful intercurrent events [6].

4. Follow-up of all participants regardless of treatment discontinuation or use of rescue treatments: Following all randomized participants, including those who discontinue randomized intervention or use rescue treatments, for the full intended observation period or until death or valid withdrawal of consent ensures maximal capture of important data on safety and efficacy. Such follow-up prevents missing data and therefore ensures reliable inference with the analyzes outlined in the following paragraph. Some steps can enhance the retention of participants. The protocol should clearly distinguish treatment discontinuation from study withdrawal, specifying that the only reason for terminating participant follow-up is death or valid withdrawal of consent. The informed consent procedure should stress the importance of continued follow-up for the intended duration of the study.

5. Primary analysis of trial data based on the ITT principle: The original ICH E9 guidance [7] stresses the importance of analysis that includes all participants, as randomized, following them all for study outcomes for the full intended observation period or until valid withdrawal of consent. Including these assessments in the analysis protects the integrity of randomization and, therefore, the trial's ability to enable unbiased estimates of the effect of interventions on outcomes. This is commonly referred to as an "intention-to-treat (ITT)" analysis. The Addendum

describes this analysis as using a "treatment policy" strategy for intercurrent events. Values for the endpoint of interest are collected and included in the analysis regardless of the occurrence of the intercurrent event. A composite strategy for certain intercurrent events may be reasonable in some cases, for example, when events in the composite are meaningful direct measures of how participants feel, function, or survive (see Section 3.5). This approach avoids using post-randomization outcomes to exclude information from randomized participants. When combined with the other design, conduct, and analysis elements outlined in this section, an ITT analysis preserves the integrity of randomization and achieves two critical objectives: (1) ensuring the ability to make valid and generalizable inference regarding the safety and efficacy of the intervention relative to control by allowing any differences observed between arms to be causally attributed to the effect of the intervention and avoiding unnecessary dependence on strong, untestable assumptions, and (2) providing information relevant to real-world settings, because such analyzes are unconditional in nature, i.e., they do not condition on post-treatment circumstances that many participants may not experience. They also properly evaluate the experimental intervention within a regimen that may include clinically appropriate ancillary care and rescue therapies. Measurements remain relevant even after participants discontinue randomized intervention or use rescue treatments. We discuss this further in subsequent sections that highlight issues with the use of while on treatment, hypothetical, and principal stratum strategies for such intercurrent events.

6. Appropriate sensitivity analysis: As described in the National Research Council monograph on missing data in clinical trials and in the Addendum [1, 2], sensitivity analyzes that systematically and comprehensively investigate and quantify the impact of missing data on inference, such as from withdrawal of consent, are important to understanding the robustness of conclusions. Comprehensive evaluation of the plausible range of assumptions regarding missing data is enhanced by including a broad spectrum of conjectures about mechanisms of non-randomly missing data.

Ensuring proper design and conduct, and then achieving proper analysis of registrational trials, is crucially important, as advocated by Scharfstein [8] in a discussion of the Addendum. This includes targeting estimands that are clinically relevant in that they compare treatment strategies that can be realistically implemented in clinical practice in groups of patients defined by measurable characteristics, as recognized by Hernán and Scharfstein [9] in their discussion of the Addendum. Unfortunately, design deficiencies often complicate the ability of trials to address clinically relevant estimands without strong and unverifiable assumptions. One example of such deficiencies is the attempt to understand the efficacy of a fixed dose or schedule of an intervention even though the trial had a titration design because the study was performed without the benefit of earlier phase trials that provided adequate data on the proper choice of dose and schedule. Another example is establishing the effect of an intervention in a hypothetical setting where rescue treatment would not be provided, even though the standard of care in the selected clinical sites permits rescue treatment. A third example comes from open-label trials that have a substantial lag between randomization and treatment initiation, resulting in a considerable number of patients randomized but not treated. Often, randomizing as proximal as possible to the time of initiating randomized interventions could have prevented that lag.

Another example of a design flaw where no estimand strategy could effectively address clinically relevant questions of central importance is a protocol crossing control patients into the experimental intervention at the time of disease progression, even though the experimental regimen is not established standard-of-care in that post-progression setting, and even though insights are important regarding effects on longer-term endpoints, such as overall survival. Many trials in oncology use this design, where investigators' reasoning seems to be that the assessment of time to disease progression is not adversely impacted by the cross-in at progression. Investigators often believe that this promise to potential participants increases enrollment rates. Such a cross-in feature may facilitate delivering a misleading presentation to potential trial participants regarding what is actually known about the benefits and risks of the experimental intervention. This practice is particularly concerning in the setting of indolent cancers. An example is the class of PI3K-inhibitors, discussed at the April 21, 2022, FDA Oncology Drugs Advisory Committee meeting; although several pivotal trials yielded strong evidence of beneficial effects on intermediate clinical outcomes, evidence eventually emerged suggesting unfavorable effects on overall survival. In many oncology trials, the systematic cross-in at disease progression has created substantive challenges for the scientific and regulatory community in making timely and reliable assessments of benefits and risks.

## 3 | Insights About Estimands and Mishandling Intercurrent Events

### 3.1 | Overview of Estimand Strategies

The Addendum states, "When defining a treatment effect of interest, it is important to ensure that the definition identifies an effect because of treatment and not because of potential confounders such as differences in duration of observation or patient characteristics." In this section, we discuss how, in many cases, estimands that apply "while on treatment", "hypothetical", or "principal stratum" strategies use post-randomization information to exclude information from randomized participants, thus failing to preserve the integrity of randomization. Consequently, observed differences may be due to confounders rather than to an effect of the treatment on a meaningful outcome. We also discuss how estimands using such strategies have the potential to mislead physicians and patients about the most important effects of an intervention. Therefore, such strategies are often problematic for use in primary analyzes, although they may be of interest in exploratory analyzes. We point out, in Section 4.3, some settings where they address clinically relevant questions under plausible assumptions.

### 3.2 | While on Treatment Strategies

The Addendum describes the potential motivation for the "while on treatment" strategy as being in settings where "*response to* *treatment prior to the occurrence of the intercurrent event is of interest . . . Events such as discontinuation of treatment, switching between treatments, or use of an additional medication may render the later measurements of the variable irrelevant or difficult to interpret even when they can be collected*." At face value, "while on treatment" analyzes may appear attractive. Researchers, physicians, and even some statisticians perceive "while on treatment" analyzes as providing valid assessments of temporal association between concurrent exposure to randomized treatment and an outcome of interest. A common argument is a discontinued intervention is unlikely to cause events occurring $X$ days after cessation where $X$ depends on biological plausibility. Further, the introduction of additional therapies after stopping randomized treatment may appear to muddy inference about the treatment under investigation. Such arguments, however, have important limitations [10, 11].

In randomized controlled trials where the experimental intervention is part of a regimen allowing standard supportive care, measurements after the participant discontinues the randomized intervention or initiates additional treatments, or both, remain clinically relevant. The intended beneficial effects, or unintended harmful effects, of the experimental intervention may influence outcomes occurring after treatment because of carry-over effects or effects on the disease process induced while on treatment, as well as effects on a participant's eligibility to receive a supportive care and the impact that such supportive care would have.

Additionally, the intercurrent events themselves, including treatment discontinuation, or initiation of rescue treatments or death, are often strongly influenced by the intervention and strongly associated with outcome. Hence, truncating follow-up at treatment discontinuation could induce substantial informative missingness and, consequently, serious inferential bias.

The "while on treatment" strategy (or perhaps more accurately, a "while-alive" strategy) has been discussed for handling death in a setting of an intervention intended to treat symptoms in patients with a terminal illness. The motivation might be to evaluate the treatment effect on symptoms before death; because such a strategy would not capture any unintended effects of the intervention on mortality, it may produce misleading information about the net effects of the intervention.

When choosing a strategy for the primary analysis, such as "while on treatment" or "while alive", having a well-defined and reliably estimated estimand is not sufficient. Ensuring inference on this estimand is meaningful and relevant to patients and prescribers is necessary. The estimand must not have a substantial risk for being misleading. For illustration, consider a trial with a clinically important "feels, functions, survives" primary endpoint. Suppose 75% of trial participants are "vigorous" and 25% are "frail". Suppose frail patients have on average less favorable outcomes but these two groups are not readily discerned by baseline characteristics. Imagine also that the experimental regimen's only causal effects are its unfavorable effects on the frail patients. Consider two settings. In the first, these effects induce frail patients to promptly both stop their therapy and experience persistent worsening outcomes on the primary endpoint. In a second setting, these unfavorable effects in frail patients lead to their acute risk of death. "While on treatment" analyzes in the first setting and "while alive" analyzes in the second both may yield misleading results, because in both situations the experimental cohort contains predominantly vigorous patients while the control cohort would include both "vigorous" and "frail" patients, and because the unfavorable effects are not captured.

An alternative to a "while alive" analysis would be an analysis addressing death through a composite strategy (see Section 3.5). One such outcome would be "days alive with quality control of symptoms over the first $X$ days after randomization". Alternatively, the full evidence from the continuous assessment of symptoms could be retained by integrating the symptom score over time the patient is alive, assigning the worst case over the interval after the patient's death. To partially disentangle effects on components of composite endpoints, supportive analysis could separately assess effects of each component.

Among issues that must be considered with the "while on treatment" estimand is whether protopathic (i.e., based on signs or symptoms truly indicative of impending events, but not yet recognized as such) or indication bias renders this estimand clinically irrelevant. For example, the experience of the EXSCEL clinical trial [12, 13], which compared major adverse cardiovascular events in patients with Type 2 diabetes treated with exenatide or placebo, suggests that restricting attention to events while participants were on the protocol-defined therapy would not capture the true effect of treatment: Patients discontinuing randomized treatment had approximately a 5- to 10-fold higher incidence rate of MACE in the 30 days following discontinuation compared to participants remaining on their assigned treatment. While it is often presumed that similar magnitudes of elevated risk following changes of therapy on each treatment arm would not greatly affect the measures of relative risk across treatment arms, the possibility of differential causes of discontinuation must also be considered. In this example, when compared to patients remaining on exenatide, a 10-fold higher rate of MACE in an average 1.5 years of follow-up was observed in the participants discontinuing blinded exenatide and later starting open-label treatment with exenatide or another glucagon-like peptide 1 receptor agonist (GLP 1 RA) in the same class. Similar comparisons within the placebo arm found only a 7-fold higher risk in patients after discontinuing placebo to start open label exenatide or another GLP 1 RA, when compared to patients who remained on placebo. Such differential effects are quite plausible: Placebo participants who experience inadequate control of hyperglycemia or worsening cardiovascular disease might add other therapies. Exenatide participants would on average have better control of hyperglycemia, so that progressing to alternative therapies might be more closely related to cardiovascular disease. Disentangling the factors leading to post-randomization changes in therapy, and their impact on biasing estimates of treatment effect, can generally not be done in a scientifically rigorous manner.

These issues would be better addressed by following all randomized subjects for study outcomes regardless of the discontinuation of randomized intervention or the use of rescue treatments and conducting an ITT analysis including all these assessments. Since intercurrent events likely would be causally affected by the experimental intervention, supportive analyzes comparing treatment arms with respect to the probability of the occurrence of such intercurrent events could provide relevant additional insights.

### 3.3 | Hypothetical Strategies

The Addendum, in describing the potential motivation for the hypothetical strategy, states: "*It may be of clinical or regulatory importance to consider the effect of a treatment under different conditions from those of the trial that can be carried out. Specifically, when additional medication must be made available for ethical reasons, a treatment effect of interest might concern the outcomes if the additional medication was not available*." Some have applied the hypothetical strategy to recover the so-called "true" or "undiluted" estimated treatment effect where an intercurrent event, such as the provision of additional non-randomized therapy, is feared to have obscured the effectiveness of the randomized treatment. If, however, an additional medication must be made available for ethical reasons, such as access to liver transplantation in patients with advanced primary biliary cholangitis, it is unclear the importance of estimating effects in a hypothetical setting where such additional medication would not be administered. Alternatively, suppose an intercurrent event resulted in cessation of randomized treatment, leading to a claim that the resulting treatment effect is diluted and underestimated. Advocates of the "hypothetical" strategy for treatment discontinuation often claim it produces an estimate of the "true", perhaps maximal, effect of uninterrupted randomized treatment allowing prescribers to advise fully adherent participants of the treatment's "true" efficacy [14, 15].

The rationale for the "hypothetical" trategy has important limitations because we can never untangle the interplay between randomized treatment, the intercurrent event, and outcome. Use of a hypothetical estimand leads to estimating effects in counterfactual settings, often requiring reliance on strong, untestable assumptions. For example, in considering a hypothetical strategy for treatment discontinuation, pre-specified analyzes often simply compare outcomes in the subsets of data defined by that post-randomization intercurrent event, using imputations based on while on treatment measures of outcome and missing at random assumptions (e.g., mixed model repeated measures, proportional hazards regression). In practice, however, the estimand should distinguish between a hypothetical population of participants who are fully adherent because they are forced to take the treatment regardless of its ill effects and an idealized hypothetical population in which no intercurrent event leading to discontinuation will occur. The handling of missing data under these variations of a hypothetical estimand are very different. With either variation, any attempted inference risks considerable bias in estimating the true treatment effect in that hypothetical setting. Furthermore, the relevance of such hypothetical scenarios is unclear, given the expected occurrence of the intercurrent events in the real world. Depending on the nature of the intercurrent event, the latter setting might be better regarded as restricting the study population to a principal stratum of patients whose tolerance of, seeming response to, and adherence with the treatment regimen would lend itself to continued treatment, and analyzes would target the prediction of principal stratum membership (see Section 3.4).

An example of limitations with a hypothetical strategy occurs in trials evaluating immunoglobulin A nephropathy (IgAN) whereby IgA protein accumulates in the kidneys, impairing long-term renal function [16]. The recent approval of SGLT2 inhibitors as effective therapy in IgAN has resulted in concerns that randomized trials comparing novel interventions with placebo in the background of established standard of care may be impacted by physicians prescribing SGLT2 inhibitors during the trial. This has led investigators to execute a hypothetical strategy, censoring trial participants at the time of SGLT2 administration or excluding those patients from the comparison of randomized treatments. Such censoring, however, may be informative and may induce bias due to potential systematic differences between arms in the types and prognoses of patients who initiate SGLT2 inhibitors. While statistical methods are available to account for informative censoring, they rely on strong and untestable assumptions, which are problematic, especially for primary analyzes in confirmatory trials. Moreover, it is not clear why an understanding of treatment effects in the hypothetical scenario where SGLT2 inhibitors are not available is relevant to real-world use. See Carroll [17] for similar concerns regarding informative censoring and the scope for serious bias in oncology trials.

Although treatment effects in hypothetical settings usually are not relevant to actual populations, they may be of interest if there were a strong reason to believe they may represent future real-world conditions, such as after the end of a war or a pandemic, and if the estimand can be estimated reliably. Examples and further discussion are provided in Section 4.3. On the other hand, in a clinical trial evaluating an immunosuppressive intervention for a hematologic malignancy, truncating follow-up of participants when hospitalized with COVID-19 infection would be problematic because the experimental intervention could affect the risks of such events. It would be implausible to achieve unbiased imputation in the presence of such informative censoring.

Estimands using a "hypothetical" strategy for intercurrent events are often proposed for trials conducted with placebo control arms in which standard-of-care approved safe and effective therapies are withheld and provided only as post-randomization rescue treatment. The argument is that, in settings where a substantial amount of rescue use is expected, an analysis using the "treatment policy" strategy for rescue use may require a sponsor to establish superiority of the experimental intervention over an active control. This concern, however, can be mitigated by alternative designs, as discussed in Section 2, that use standard-of-care control arms. An alternative approach to understanding the effects of an experimental intervention in settings where another medication would not be used is to conduct a trial in a clinical setting where, under available standard-of-care, use of that additional medication would be unlikely prior to the assessment of the primary endpoint. For example, in alcoholic hepatitis, unlike in primary biliary cholangitis, many clinical sites do not provide liver transplantation as standard-of-care, so a clinical trial in alcoholic hepatitis with a 90-day survival endpoint could be conducted in such sites if there were interest in the efficacy of an experimental intervention in the absence of liver transplantation. This would enable us to use an ITT analysis to address that clinical question of interest. If this approach would lead to conducting such trials in lower-to-middle-income countries, then, by the Distributive Justice principle [18], a viable plan should be in place to enable the participating population to receive that experimental intervention, if subsequently shown safe and effective.

### 3.4 | Principal Stratum Strategies

The Addendum describes the potential motivation for the "principal stratum" strategy as follows: "*Certain clinical events can also be intercurrent events, when their occurrence, or non-occurrence, defines a principal stratum of interest. Examples include tumor shrinkage defining objective response when assessing treatment effect on duration of response in oncology.*" The Addendum refers to principal stratification as being based on potential outcomes, e.g., based on subgroups of participants who would have intercurrent events had they been assigned to a specific treatment arm.

Extensive literature is available on methods that might best estimate effects among the clinical trial participants in the principal strata of interest [19]. However, the principal strata usually cannot be fully identified, and such analyzes rely on strong and unverifiable assumptions and have a high risk of bias, especially if the intervention could affect the occurrence of the intercurrent event of interest. In addition, analyzes described as addressing an estimand using a principal stratum strategy for an intercurrent event often simply compare outcomes in the post-randomization subgroups defined by that intercurrent event (e.g., in evaluating duration of response in oncology trials). Based on this implementation, the "principal stratum" estimand focuses on an "improper subgroup" of participants, i.e., a subgroup only known after a given intercurrent event occurs [20], and the Addendum wisely distinguishes between such analyzes and analyzes based on principal strata defined by the potential for intercurrent events. Furthermore, such analyzes are of questionable relevance to the prescriber and patient because the stratum in which the patient would fall is unknowable at the time of treatment decisions.

The potential for bias is illustrated through the typical assessment of duration of response in oncology trials by conditioning only on participants experiencing tumor response. As noted in CHMP guidance [21], analyzes based only on responding patients are problematic because such groups are non-randomized and may not be comparable; therefore, they should not be subject to formal statistical inference. An analysis of "time in response" avoids such bias if it assigns non-responders a zero duration of response [5, 22]. This analysis also could provide increased sensitivity. For example, a doubling in objective response rate with a doubling in duration of response (calculated using only responders) translates to a true four-fold increase in time in response [5].

The Addendum describes another example that potentially motivates the "principal stratum" strategy, stating, "*It might be desired to know a treatment effect on severity of infections in the principal stratum of patients becoming infected after vaccination.*" Because randomization occurs before the existence of the clinical condition, (in this instance, infection post-vaccination), and because its occurrence cannot be reliably predicted at baseline, to obtain interpretable and unbiased results, an ITT analysis in all randomized participants, rather than a "principal stratum" approach, would address the clinically important question about treatment effect on risk of infection, (i.e., vaccine efficacy), and, separately, regarding risk of infections leading to severe disease, (as has been done in the evaluation of vaccines for malaria [23] and for COVID-19 [24]).

### 3.5 | Composite Strategies

Incorporating intercurrent events of compelling clinical relevance into a composite primary endpoint, (i.e., using a composite strategy), may sometimes be desirable. For example, a participant who has died has had a bad outcome; the definition of the endpoint should reflect this. A composite endpoint including death enables analyzes based on the ITT principle. In a setting assessing time to a clinically important event, the composite outcome measure that includes death, "event-free survival", would be a proper clinical endpoint assessable using an ITT analysis. Other examples of compelling events that a composite outcome could address might include the surgical removal of the nasal polyp in nasal polyposis, leg amputation when assessing symptoms of diabetic foot ulcers, knee replacement in osteoarthritis of the knee, and organ transplant in various conditions. Similarly, in cardiovascular (CV) outcome trials where interest may lie in CV death, other events such as nonfatal myocardial infarction, nonfatal stroke, and non-CV deaths represent compelling outcomes that are frequently included in a composite outcome of major CV events.

In oncology, using the composite endpoint, "progression-free-survival," rather than "time to progression" both enhances the clinical relevance of the measure and enables use of ITT analyzes. An illustration of this concept in the setting of a treatment in a hospitalized or ICU population would be to account for death by using "days alive and out of the hospital" or "days alive and out of the ICU" as the endpoint. These would be proper measures, unlike "days in the hospital" or "days in the ICU".

While the Addendum states, "*a patient who discontinues treatment because of toxicity may be considered not to have been successfully treated*," considering treatment discontinuation to be an event in a composite endpoint would meaningfully weaken the clinical relevance of that endpoint, in part because participants could experience considerable benefit from a therapy even after they discontinue its administration. Furthermore, better adherence rather than true clinical benefit could lead to an observed favorable effect on a composite outcome that defines adherence as a component.

An important issue with the Composite Strategy is the inherent challenge in disentangling treatment effects on components [25, 26]. As mentioned previously, to address this, one could assess effects on components individually in supportive analyzes [27], using treatment policy strategies when possible and non-treatment policy strategies when not.

## 4 | Additional Recommendations

### 4.1 | Non-inferiority Trials

The Addendum indicates, "*The considerations informing the construction of the estimand to support regulatory decision making based on a non-inferiority or equivalence objective may differ to those for the choice of estimand for a superiority objective.*" While the importance of ensuring the integrity of non-inferiority trials motivates efforts to avoid irregularities in quality of trial conduct, the advantages of the trial elements discussed in Section 2, including the use of ITT analyzes, also apply to non-inferiority trials. The Addendum properly recognizes "*the importance of minimizing the number of protocol violations and deviations, non-adherence and study withdrawals*" in non-inferiority trials. As Fleming et al. [28] indicate, "*The preferred approach to enhancing the integrity and interpretability of the non-inferiority trial should be to establish performance standards for measures of quality of trial conduct (e.g., targets for enrollment and eligibility rates, event rate, adherence and retention rates, cross-in rates, and currentness of data capture) when designing the trial, and then to provide careful oversight during the trial to ensure these standards are met, with the "as randomized" analysis being the primary analysis . . .*".

### 4.2 | Addressing Prevention and Treatment of Missingness With Appropriate Assumptions and Sensitivity analyzes

The Addendum asserts, "*To reduce missing data, measures can be implemented to retain subjects in the trial.*" This statement is important. As Fleming [29] indicates, "*The reliability and interpretability of clinical trials can be substantially reduced by missing data . . . Although rational imputation methods may be useful to treat missing data, these methods depend on untestable assumptions . . . The preferred and often only satisfactory approach to addressing missing data is to prevent it.*" Procedures should be in place to maximize the likelihood of obtaining outcome data at scheduled times of evaluation for all surviving participants who have not withdrawn consent [30]. Doing so enables implementation of ITT analyzes. Despite best efforts at retention, if some data are missing, assumptions about missing data, (e.g., in an imputation model), should be "centered" around the best projections for the participants' outcomes had they been captured and should account for the corresponding uncertainties around those projections, (e.g., with multiple imputation). This may involve imputation methods based on the participants' treatment group, specific insights about the reason for their non-retention, the level of their clinical response at the time of discontinued follow-up, and the likelihood their treatment benefits would persist. In addition, sensitivity analyzes such as tipping point analyzes should be conducted to evaluate systematically and comprehensively whether the conclusions hold up under plausible violations in the assumptions about missing data, with the recognition that the mechanisms of missingness can differ by treatment arm.

### 4.3 | Exceptions Where Alternative Approaches May be Justified

In rare settings, an alternative estimand approach, such as the use of a hypothetical strategy for handling certain intercurrent events, may be justified if it addresses a clinically relevant question with plausible assumptions. For example, in some settings, compelling arguments justify excluding post-randomization information without bias and without compromising the interpretability of trial results. Such a situation could arise, as in the examples below, when these exclusions are clearly independent of both the effect of the randomized interventions and knowledge of participant randomized assignment. One illustration would include the truncation of follow-up post February 2022 in Ukrainian sites due to the war in Ukraine. This would address a relevant hypothetical question (to wit, what would be the effect of treatment if there were no war?), because a geo-political war meaningfully impacts access and adherence to treatment and evaluations of outcomes in a manner meaningfully inconsistent with realistic, future non-war clinical settings.

A second illustration comes from HIV-prevention trials. Exclusion of all participant follow-up during intervals in calendar time, following pre-specified criteria, may be justified when the presence of peak levels of COVID-19 waves would meaningfully adversely impact treatment adherence as well as assessment of outcomes in a manner meaningfully inconsistent with realistic, future clinical settings in the absence of a global pandemic.

As a third illustration, one could consider the exclusion of the small number of participants randomized but never treated in a blinded trial, when neither the effect nor knowledge of treatment assignment could impact their exclusion. The FLO-ELA open-label trial provides a potentially acceptable extension of this idea. The trial compared two methods of fluid delivery for patients following surgery [31]. While the preferred design approach would be to randomize participants as proximal as possible to the time of initiation of randomized interventions and then to perform a per randomization analysis, randomization in the FLO-ETA trial needed to be performed just prior to surgery rather than nearer to the initiation of post-surgical fluid delivery because the investigators recognized that "each fluid delivery method takes substantial preparation". While this could result in occurrence of randomized but non-surgically treated patients, the exclusion of such patients from the analysis could be justified due to the implausibility that the intercurrent event of surgery cancelation could be related to randomized assignment, as well as the recognition that the number of exclusions would be ≤ 2% (as confirmed in Supplemental File 1 of the published protocol), because randomization typically would not occur until "the patient arrives in the theater site for surgery".

## 5 | Conclusions

The Addendum has stimulated cross-disciplinary discussions on how to improve clinical trials. Its careful discussion of the importance of matching the estimator to the estimand has focused trialists' attention on how best to think about ensuring that the trial's methodology meets its aims. It has positively influenced the selection of the trial population, primary endpoint, and population-level summary measure for comparing treatment arms. It also recognizes the critically important role of randomization and the desirability of including all randomized participants in the analysis; it emphasizes the need for reducing missing data and urges prospectively planning how to address intercurrent events such as the discontinuation of randomized treatment and the initiation of rescue treatments. It raises awareness about the risk of bias when subgroup analysis or termination of follow-up is based on the occurrence of an intercurrent event, recognizes that results under hypothetical conditions might not be of clinical or regulatory interest, and asserts that the assumptions underlying the analysis should be justifiable, plausible, and supported by appropriate sensitivity analyzes.

Unfortunately, many clinical trial protocols and statistical analysis plans have used strategies for handling intercurrent events described in the Addendum in ways that fail to preserve the integrity of randomization or that do not capture meaningful effects of the intervention. Implementation of the estimand framework using “while on treatment”, “hypothetical”, or “principal stratum” strategies for handling intercurrent events is problematic when serving as the trial’s primary confirmatory analysis regarding causal effects of interventions rather than in their more appropriate role as a supportive descriptive analysis. In many cases, these estimand strategies fail to preserve the integrity of randomization because they use post-randomization outcomes to exclude information from randomized participants. Failure to preserve the integrity of randomization can lead to non-comparable groups and inherent bias in analyzes, in turn preventing attribution of causality to treatment. Such analyzes also often lack generalizability and have the potential to mislead prescribing physicians and participants about the most important effects of the intervention.

In contrast, there are critically important advantages of properly designed randomized clinical trials with standard-of-care control arms and with primary endpoints that are measures of how a patient “feels, functions, survives” or are validated surrogates, and that are properly conducted with high levels of retention. Such trials reduce incentives to implement estimand strategies that fail to preserve the integrity of randomization and depend on untestable assumptions. In turn, such trials also enable obtaining valid and unbiased inference about issues of central clinical relevance when applying a treatment policy estimand strategy for handling intercurrent events such as treatment discontinuation and rescue medication use, while potentially also incorporating other meaningful intercurrent events, such as death, into the primary endpoint applying a composite strategy. Such estimands provide meaningful information relevant to real-world settings, as they are unconditional in nature, i.e., they do not condition on post-treatment circumstances that may not be the experience of many participants, and properly evaluate the experimental intervention within a regimen that includes possible ancillary care and rescue therapies that would be clinically appropriate in a real-world setting.

In conclusion, ensuring proper design, conduct, and analysis of registrational trials is necessary to achieve valid inference from them about clinically important hypotheses. Without proper design, conduct, and analysis, no meaningful estimand can be reliably evaluated. In randomized trials, this should be done in a manner that incentivizes and enables appropriate use of the treatment policy strategy, which preserves the integrity of randomization and enables reliable evaluation of intervention effects on clinically meaningful endpoints. This is essential to answering the questions of greatest importance in the pursuit of more effectively preventing and treating diseases.

### Acknowledgments

The authors thank the reviewers, the associate editor, and the editor for their valuable comments and suggestions that improved this article.

### Conflicts of Interest

The authors declare no conflicts of interest.

### Data Availability Statement

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

### References

1. European Medicines Agency, *ICH E9 (R1) Addendum on Estimands and Sensitivity Analysis in Clinical Trials to the Guideline on Statistical Principles for Clinical Trials* (ICH, 2020).

2. National Research Council (US) Panel on Handling Missing Data in Clinical Trials, *The Prevention and Treatment of Missing Data in Clinical Trials* (National Academies Press, 2010), 162.

3. L. M. Friedman, C. D. Furberg, D. L. DeMets, D. M. Reboussin, and C. B. Granger, *Fundamentals of Clinical Trials*, 5th ed. (Springer, 2015), 550.

4. R. Rothwell, N. P. Nikolov, J. W. Maynard, and G. Levin, “Noninferiority Trials to Evaluate Drug Effects in Rheumatoid Arthritis,” *Arthritis & Rhematology* 72, no. 8 (2020): 1258–1265.

5. D. L. DeMets, B. M. Psaty, and T. R. Fleming, “When Can Intermediate Outcomes Be Used as Surrogate Outcomes?,” *Jama* 323, no. 12 (2020): 1184–1185.

6. A. J. Sankoh, H. Li, and R. B. D’Agostino, Sr., “Use of Composite Endpoints in Clinical Trials,” *Statistics in Medicine* 33, no. 27 (2014): 4709–4714, https://doi.org/10.1002/sim.6205.

7. European Medicines Agency, *ICH E9 Statistical Principles for Clinical Trials—Scientific Guideline* (European Medicines Agency, 1998).

8. D. O. Scharfstein, “A Constructive Critique of the Draft ICH E9 Addendum,” *Clinical Trials* 16, no. 4 (2019): 375–380.

9. M. A. Hernán and D. Scharfstein, “Cautions as Regulators Move to End Exclusive Reliance on Intention to Treat,” *Annals of Internal Medicine* 168, no. 7 (2018): 515–516.

10. F. Yang, J. Wittes, and B. Pitt, “Beware of On-Treatment Safety Analyses,” *Clinical Trials* 16, no. 1 (2019): 63–70.

11. T. R. Fleming, M. D. Rothmann, and H. L. Lu, “Issues in Using Progression-Free Survival When Evaluating Oncology Products,” *Journal of Clinical Oncology* 27, no. 17 (2009): 2874–2880.

12. R. R. Holman, R. J. Mentz, V. P. Thompson, et al., “Effects of Once-Weekly Exenatide on Cardiovascular Outcomes in Type 2 Diabetes,” *New England Journal of Medicine* 377, no. 13 (2017): 1228–1239, https://doi.org/10.1056/NEJMoa1612917.

13. M. A. Bethel, S. R. Stevens, J. B. Buse, et al., “Exploring the Possible Impact of Unbalanced Open-Label Drop-In of Glucose-Lowering Medications on EXSCEL Outcomes,” *Circulation* 141 (2020): 1360–1370.

14. O. N. Keene, H. Lynggaard, S. Englert, V. Lanius, and D. Wright, “Why Estimands Are Needed to Define Treatment Effects in Clinical Trials,” *BMC Medicine* 21, no. 1 (2023): 276.

15. O. N. Keene, D. Wright, A. Phillips, and M. Wright, “Why ITT Analysis Is Not Always the Answer for Estimating Treatment Effects in Clinical Trials,” *Contemporary Clinical Trials* 108 (2021): 106494, https://doi.org/10.1016/j.cct.2021.106494.

16. D. V. Rizk, B. H. Rovin, H. Zhang, et al., “Targeting the Alternative Complement Pathway With Iptacopan to Treat IgA Nephropathy: Design and Rationale of the APPLAUSE-IgAN Study,” *Kidney International Reports* 8, no. 5 (2023): 968–979.

17. K. J. Carroll, “Analysis of Progression-Free Survival in Oncology Trials: Some Common Statistical Issues,” *Pharmaceutical Statistics* 6, no. 2 (2007): 99–113.

18. J. Lamont and F. Favor, "Distributive Justice," in *Stanford Encyclopedia of Philosophy* (Metaphysics Research Lab, Philosophy Department, Stanford University, 2017).

19. I. Lipkovich, B. Ratitch, Y. Qu, X. Zhang, M. Shan, and C. Mallinckrodt, "Using Principal Stratification in Analysis of Clinical Trials," *Statistics in Medicine* 41, no. 19 (2022): 3837–3877, https://doi.org/10.1002/sim.9439.

20. S. Yusuf, J. Wittes, J. Probstfield, and H. A. Tyroler, "Analysis and Interpretation of Treatment Effects in Subgroups of Patients in Randomized Clinical Trials," *Jama* 266, no. 1 (1991): 93–98.

21. European Medicines Agency, *Guideline on the Evaluation of Anticancer Medicinal Products in Man* (European Medicines Agency, 2017).

22. S. H. Ellis, K. J. Carroll, and K. Pemberton, "Analysis of Duration of Response in Oncology Trials," *Contemporary Clinical Trials* 29, no. 4 (2008): 456–465.

23. RTS, S Clinical Trials Partnership, "Efficacy and Safety of RTS,S/AS01 Malaria Vaccine With or Without a Booster Dose in Infants and Children in Africa: Final Results of a Phase 3, Individually Randomized, Controlled Trial," *Lancet* 386, no. 9988 (2015): 31–45. Erratum in: Lancet. 2015; 386 (9988): 30.

24. WHO Ad Hoc Expert Group on the Next Steps for Covid-19 Vaccine Evaluation, "Placebo-Controlled Trials of COVID-19 Vaccines–Why We Still Need Them," *New England Journal of Medicine* 384, no. 2 (2021): e2.

25. G. Cordoba, L. Schwartz, S. Woloshin, H. Bae, and P. C. Gøtzsche, "Definition, Reporting, and Interpretation of Composite Outcomes in Clinical Trials: Systematic Review," *BMJ* 341 (2010): c3920.

26. S. J. Pocock, C. A. Ariti, T. J. Collier, and D. Wang, "The Win Ratio: A New Approach to the Analysis of Composite Endpoints in Clinical Trials Based on Clinical Priorities," *European Heart Journal* 33, no. 2 (2012): 176–182.

27. G. Y. Chi, "Some Issues With Composite Endpoints in Clinical Trials," *Fundamental & Clinical Pharmacology* 19, no. 6 (2005): 609–619, https://doi.org/10.1111/j.1472-8206.2005.00370.x.

28. T. R. Fleming, K. Odem-Davis, M. Rothmann, and Y. Li Shen, "Some Essential Considerations in the Design and Conduct of Non-inferiority Trials," *Clinical Trials* 8, no. 4 (2011): 432–439, https://doi.org/10.1177/1740774511410994.

29. T. R. Fleming, "Addressing Missing Data in Clinical Trials," *Annals of Internal Medicine* 154, no. 2 (2011): 113–117.

30. J. Wittes, "Missing in Action: Preventing Missing Outcome Data in Randomized Clinical Trials," *Journal of Biopharmaceutical Statistics* 19, no. 6 (2009): 957–968.

31. M. R. Edward, G. Forbes, N. Walker, et al., "Fluid Optimisation in Emergency Laparotomy (FLO-ELA) Trial: Study Protocol for a Multi-Centre Randomised Trial of Cardiac Output-Guided Fluid Therapy Compared to Usual Care in Patients Undergoing Major Emergency Gastrointestinal Surgery," *Trials* 24, no. 1 (2023): 313, https://doi.org/10.1186/s13063-023-07275-3.