Framing time-to-event estimands and censoring mechanisms in oncology in light of the estimands framework

Jonathan Siegel (Bayer), Lynda Grinsted(AstraZeneca), Feng Liu (Fate Therapeutics), Hans-Jochen Weber (Novartis), Stefan Englert (AbbVie), Michelle Casey (Pfizer) & on behalf of the Industry Working Group on Estimands in Oncology

Short Title: Censoring and censoring mechanisms in oncology with estimands

Keywords: Estimands, Oncology, Survival Analysis, Clinical trials, Censoring

Abstract

In oncology clinical trials with time-to-event endpoints, censoring rules have traditionally been defined and applied following standard approaches based on longstanding regulatory guidelines. The estimand framework (addendum to the ICH E9 guideline) calls for precisely defining the treatment effect of interest to align with the clinical question of interest and requires predefining the handling of intercurrent events that occur after treatment initiation and either preclude the observation of an event of interest or impact the interpretation of the treatment effect. In the context of time to event endpoints, this requires a careful discussion on how censoring rules are applied. We discuss a practical problem in clinical trial design and execution, i.e. in some clinical contexts it is not feasible to systematically follow patients to an event of interest. We discuss what censoring means in such contexts and alternative strategies available to address it. We introduce terminology to distinguish types of censoring. We provide recommendations for trial design, stressing the need for close alignment of the clinical question of interest and study design, impact on data collection and other practical implications. We discuss the use of sensitivity and supplementary analyses to examine such censoring assumptions.

1. Introduction

With the release of the addendum to the ICH E9 guidance\(^1\), the estimand framework is being applied widely in the design and analysis of clinical trials. The estimand framework provides terminology to clearly describe what a clinical trial is intending to assess and consequently, which data need to be collected and how these data are to be analysed. In light of the estimand framework, this paper provides a discussion on the impacts of censoring in time to event analyses on the clinical question of interest. This paper is a product of the Pharmaceutical Industry Working Group on Estimands in Oncology’s Censoring Mechanisms Subteam.

For an overview of the estimand framework generally, see e.g. Akacha et al.\(^2\); Ratich et al.\(^3\); and Malinckrodt et al.\(^4\) and for estimands in time-to-event endpoints, see Rufibach\(^5\)
In oncology trials, assessing the treatment effect by time to event endpoints is typically of high interest and referenced in health authority guidelines. In the past, censoring rules for time to event endpoints have sometimes been defined within protocols and analysis plans without giving consideration to the fact that different event and censoring rules might address different clinical questions of interest.

In this paper we consider that while it is often desirable to follow all patients to the clinical event of interest, in some cases it is not possible or not feasible to follow patients through and beyond particular intercurrent events. Lack of follow-up in such cases requires careful definition and interpretation of the relevant estimand. We identify a number of real-world clinical situations where consistent follow-up may not be feasible and introduce strategies to address it.

In addition to attempting to address the inability to follow patients until the clinical event of interest, we provide general strategies and recommendations for clarifying the scientific question of interest, defining estimands, selecting strategies for addressing intercurrent events, collecting data, censoring, and defining sensitivity analyses in the context of time-to-event endpoints. We also discuss more general cases where the originally desired estimand might be infeasible in a particular context, and discuss ways to develop alternatives.

2. Background

Classical time-to-event analysis methods are based on the critical assumption that when the event of interest has not yet occurred and patients are censored, this censoring is non-informative with respect to the scientific question of interest and hence can be appropriately handled by censoring. Non-informative censoring is independent of, and provides no information about, the question of interest.

In contrast, patients who discontinue participation in a trial for treatment-related reasons such as perceived lack of efficacy or treatment side effects may well have a different risk of experiencing the event in the future than other patients (informative censoring). It has traditionally been common to handle certain types of intercurrent events, such as subsequent therapy, with censoring even if the event of interest is observed afterwards.

A recurring problem in oncology clinical trials with survival estimands is situations where all or a substantial fraction of patients cannot feasibly be followed beyond an intercurrent event when follow-up is mandated by the chosen strategy. Examples include:

- Patients can never be followed beyond terminal events, such as death. As mortality is, unfortunately, pervasive in cancer, appropriately accounting for death requires careful consideration in virtually all estimands in an oncology trial.
- Within some open-label trials or trials with functional unblinding due to side effects, large fractions of patients randomized to placebo have left the trial shortly after randomization. An example is the Checkmate-037 trial, where 37% of patients
randomized to placebo immediately went off treatment and took other immunotherapies as subsequent therapy. See generally Manitz (2021).

- Although it is increasingly possible to continue clinic visits and assessments beyond treatment discontinuation, it is not always possible to do so beyond events such as radiological progression. Patients might for example move to another trial. Although devices for recording Patient Reported Outcomes, patient diaries, etc. are increasing the scope of home data collection, many kinds of assessments can still only be collected reliably at clinic visits. When clinic visits cease, assessments dependent on clinic visits also cease.

Every clinical trial has a unique scientific question of interest in the context of a specific treatment condition, population, variable, intercurrent-event strategy, and population-level summary measure. Typical clinical events in such a setting that have led to censoring need to be reviewed and discussed with regard to their relationship to the clinical question. Bias can only be determined in the context of a specific clinical question. An approach that is biased for addressing one kind of clinical question may not be biased for another. This perspective underlines the need to define study-specific rules tailored to the scientific question of interest and to consider the necessary data collection and patient follow-up to address the scientific question of interest. This is particularly important when considering cases where a substantial amount of censoring due to intercurrent events may occur.

3. **Time to event endpoints in oncology**

Substantial progress has been made over recent years to understand, diagnose and to treat cancer. Life expectancy has increased in many indications, however, prolonging life remains the key objective in cancer patients. Therefore, overall survival is the gold standard endpoint and well acknowledged by Health Authorities. Other time to event endpoints like PFS, disease free survival (DFS) or event-free survival (EFS) have been accepted particularly when surrogacy to overall survival or clinical relevance can be demonstrated.

For all time to event endpoints, the time between a triggering event (e.g. randomization or first dose) until a clinical event of interest occurs is analysed.

Typical time to event endpoints in oncology include:

- **OS** is the time from randomization (or first dose, in non-randomized trials) until death from any cause
- **PFS** is the time from randomization (or first dose, in non-randomized trials) to objective disease progression, or death from any cause, whichever occurs first
- **DFS** is the time from randomization (or first dose, in non-randomized trials) to objective disease recurrence or death from any cause, whichever occurs first
- **EFS** is typically indication specific, but not necessarily consistent in different protocols. For example, for acute myeloid leukaemia EFS was defined as the time from randomization to (induction) treatment failure, relapse for those who have
induction treatment success (e.g. complete remission), or death from any cause, whichever occurs first.\textsuperscript{11,12}

4. Missing data and intercurrent events

During the follow-up of patients, situations might occur that either preclude the observation or affect the interpretation of the variable. These are defined in the ICH E9 addendum as intercurrent events. It is important to differentiate missing data from intercurrent events. According to the estimands guidance, missing data will not bias the estimator nor change the estimand, but weakens the strength of the estimator, typically in loss of statistical power. In contrast, intercurrent events could change the intended estimand and, depending on the estimand, could introduce bias if ignored. Thus, all potential intercurrent events require an \textit{a priori} assessment at the design stage to define the appropriate approach on how to deal with them along with a discussion how censoring is applied. According to the estimands guidance\textsuperscript{1}, administrative closure of the trial results in missing data. An individual decision to discontinue treatment based on perceived lack of efficacy or toxicity will generally represent a potential intercurrent event.

Both missing data and intercurrent events can occur while the patient is still being followed. For time-to-event variables, the observation period of a patient can be separated into the time up to the occurrence of an intercurrent event or endpoint of interest and the time thereafter. Just by the chronological sequence it can be concluded that the risk of the event of interest for the time interval up to the intercurrent event is not impacted by the intercurrent event but the time thereafter is.

A key decision in the estimands framework is whether a particular event or class of events results in missing data or should be classified as intercurrent. In order to make this classification, it is critically important to collect necessary data, particularly data about the reasons for discontinuation or loss to follow-up for the relevant assessments. Under the estimand framework, cases where loss to follow-up has no apparent relationship to the treatment effect are considered as missing data. Cases where there is a potential relationship should generally be considered intercurrent events or integrated in other attributes of the estimand. Absent a clear scientific rationale for assuming otherwise lost to follow-up should be considered as potentially related to treatment.

Given that clinical events may be informative, data collection on the outcome of interest and events that are not \textit{a priori} defined intercurrent event (but may be considered as such with growing evidence) should continue where appropriate to the estimand and feasible.

5. Strategies for handling intercurrent events

Both missing data and intercurrent events can occur without the need for censoring. An isolated missed appointment does not generally prevent subsequent follow-up and documentation of the event of interest, is often unrelated to treatment effect, and is regarded
as missing data. Many potentially biasing events – for example changes in relevant concomitant medications, – represent intercurrent events, but similarly do not result in a patient being lost to follow-up for the event of interest. In both cases, depending on strategy used, there may not be a need to censor.

Frequently observed intercurrent events in oncology trials include deaths, discontinuation of treatment due to toxicity or subjective clinical progression and initiation of new anti-cancer therapy. In the presence of intercurrent events, the assumption of non-informativity is unlikely to hold.

Applicability of the ICH E9 (R1) Estimands guidance’s strategies for addressing intercurrent events are described below. Treatment policy and hypothetical strategies generally impact the treatment definition, especially when the ICE is treatment related. Principal stratum strategies impact the population definition. Composite and while-on-treatment strategies impact the variable definition.

Different estimand strategies can be applied to intercurrent events, depending on the estimand of interest. It is important to keep in mind the estimands framework can involve more than one strategy per estimand. Within each estimand, different strategies can be and often will need to be applied to different intercurrent events within the same estimand.

**Treatment Policy Strategy**

The term “treatment policy strategy” comes from the idea of evaluating the effect of the complete sequence of treatments, beginning with randomization and including all treatments thereafter, through treatment switching. This strategy, however, is not limited to treatment switching. It can potentially be applied to any intercurrent event. When applied to a general intercurrent event, the research question evaluates all outcomes up to the event of interest, through and beyond the applicable intercurrent event. A treatment policy strategy evaluates the total effect of the originally assigned treatment, where all post-randomization effects are ignored, left unspecified, and regarded as effects of the original treatment assignment. Accordingly, a treatment policy strategy considers all data observed relevant to the scientific question of interest. For reasons explained by Fleming et al., when correctly executed, it is generally the strategy most likely to preserve the assumption that censoring is non-informative.

Because a treatment policy strategy considers the applicable intercurrent event irrelevant to the outcome and looks to what occurred after it, valid execution of a treatment policy strategy requires follow-up consistent with this approach. Accordingly, execution of a treatment policy strategy requires consistent follow-up through and beyond intercurrent events until the event of interest is observed. This degree of follow-up can be challenging to obtain in some clinical trials and may be infeasible in some cases. For example, patients may wish to withdraw from a trial following treatment discontinuation, or for reasons related to treatment efficacy or safety. Thus, informative censoring is likely unavoidable.
A distinction should be made between situations where failure to follow patients becomes pervasive and systematic, and situations where it remains relatively isolated and occasional. Where failure to follow patients is pervasive and systematic, the assumptions underlying a treatment policy strategy may be infeasible, and a strategy that does not require or assume such follow-up will occur in most patients may be more appropriate.

**Composite Strategies**

Composite strategies make intercurrent events part of the variable (or endpoint), by making the event of interest a composite event that includes both the original event and the intercurrent one. This strategy is particularly useful and provides an alternative to censoring when the intercurrent event is positively correlated with, and highly informative of, the event of interest. Perhaps the best-known example in oncology is PFS, a composition of progression and death events. Composite strategies may be appropriate in a wide variety of contexts.

**Hypothetical Strategies**

Hypothetical strategies ask what would have happened in a counterfactual scenario, e.g., if the intercurrent event had not occurred. Such a strategy may be particularly appropriate when the relevant intercurrent event only occurs because of clinical conditions and would not occur in real-world clinical practice.

Hypothetical strategies tend to remove events from the clinical question of interest. Generally speaking, events following the intercurrent events are not relevant. By censoring after an intercurrent event, a simple hypothetical strategy is applied. This traditional approach asks what the treatment effect would be if the intercurrent event had not occurred, assuming that patients censored for the intercurrent event have the same hazard of experiencing an event as those not censored.

Censoring for intercurrent events like treatment switching is often informative, and therefore, where appropriate, alternative strategies should be considered.

More sophisticated hypothetical approaches based on counterfactuals, such as inverse probability censoring weighting (IPCW) or rank preserving structural failure time (RPSFT), are also available. Counterfactual survival times are survival times that would have occurred if the intercurrent event had not happened, as estimated by a model. Considerations for these methods, and contexts when use of a hypothetical strategy might be appropriate to consider, are discussed in the next section.

**While-on-Treatment Strategy**

A while-on-treatment strategy is only concerned with what happens up to the time of the intercurrent event. In a while-on-treatment strategy, anything after the intercurrent event is
considered irrelevant to the scientific question of interest, and hence data collection need not continue past it. The strategy can be applied to any intercurrent event, not just withdrawal from study treatment.

A motivating example is palliative treatment, which is generally not assumed to prolong survival. Clinical interest lies in evaluating alleviation of symptoms prior to death, whenever death occurs. However, a while-on-treatment strategy may be appropriate for many other contexts where what happens after the intercurrent event is not considered relevant to the research question.

In the context of time-to-event estimands, there are two basic models used to implement a while-on-treatment strategy, the cause-specific hazards model, and a dependent-causes model, generally implemented through competing risk methods like the Cumulative Incidence Function and the subdistribution hazards model. If independent causes can be assumed, for example time to cause-specific death where causality attribution can be reliably established and independence is a reasonable assumption, then a while-on-treatment strategy can be implemented through a cause-specific hazards model, which uses standard censoring and hence represents an interpretation of standard proportional hazards and related censoring models and methods (e.g. Kaplan-Meier, Cox, log-rank, etc.). As discussed above, causal independence can rarely be reliably established in oncology, and time to cause-specific event estimands are accordingly not common. When the independence assumption lacks an unequivocal basis, which we suggest is often and indeed usually the case in oncology, a while-on-treatment strategy can be implemented using a competing risk method with the intercurrent event classified as a competing risk event. Although the subdistribution-hazards (Fine-Gray) model sometimes used for modelling and testing in this context has been criticized as not having a causal interpretation, the descriptive Cumulative Incidence Function (CIF) does not have this issue.

**Principal Stratum Strategy**

A principal stratum strategy attempts to study only patients in whom the intercurrent event of interest is not expected to occur, generally using a model to predict such patients from baseline characteristics. Accordingly, the strategy addresses the intercurrent event by removing patients expected to experience it from the study population. Once a principal stratum in whom the applicable intercurrent event is expected not to occur is identified, ordinary censoring can be applied to patients for whom the intercurrent event occurs despite the model’s prediction that it will not be applied. Just as too high a proportion of patients no longer being followed due to an intercurrent event can invalidate a treatment policy strategy, too high a proportion of patients experiencing the intercurrent event, while not changing the strategy, can render the model underlying a principal stratum model nonpredictive and invalid. A principal stratum strategy has not been common in regulatory oncology clinical trials.
5. **Kinds of intercurrent events**

In addition to being considered a gold standard for establishing efficacy, overall survival is the primary variable where it is most feasible to continue to follow patients past intercurrent events and hence a treatment policy strategy is often implemented.

PFS is a common alternative primary endpoint to OS given the duration of follow-up often necessary for OS and the clinical benefit associated with delaying progression. However, intercurrent events may impede follow-up until the time of progression for all participants within a clinical trial.

Approaches and recommendations for handling of typical kinds of intercurrent events within oncology trials are as follows:

**Death.** Death is the quintessential terminal event. Accordingly, a *treatment policy strategy* cannot be used for death, and another strategy should be used. The interpretation that systematic censoring for intercurrent events is a *hypothetical strategy* does not apply here, as the research question underlying a hypothetical strategy, what would have happened if the patient had not died, is not generally a meaningful scientific question. While censoring for death could potentially be interpreted as an independent-causes while-on-treatment (while-alive) strategy, this strategy would only be valid if death is truly independent of the event of interest, which in unlikely in the context of a disease with a high mortality outcome. Thus it is not generally recommended to censor for death in oncology trials absent special justification. On the other hand, a *composite strategy* is often a good choice due to the tendency for death to be highly correlated with other clinical events. Perhaps the most common example of a composite strategy involves time to progression, which has largely been replaced by the composite endpoint progression-free survival. In situations where there is only concern about the event of interest prior to death, a while on treatment (while alive) strategy that does not assume non-informativity can be implemented through a competing risk approach. An example would be the competing risk approach to time to bone fracture with death as a competing risk event.21

**Discontinuation of treatment due to toxicity.** Participants are increasingly being followed beyond treatment withdrawal, rendering a *treatment policy strategy* generally feasible. For some endpoints of interest, a *composite strategy* may be appropriate. Where patients are not followed past end of treatment, censoring at treatment discontinuation, explicit or implicit, would induce a *hypothetical strategy*. The research question involved, what time to the event of interest if patients had not experienced toxicity, could in some cases be relevant. However, it is often not reasonable to assume that early discontinuation from treatment has no effect on future prognosis. In cases where it is inappropriate or infeasible to follow patients beyond treatment withdrawal, either the real scientific question of interest may not be concerned with what happens afterwards, or not being concerned with what happens afterwards may represent a compromise scientific question that can feasibly be addressed
under the circumstances. In this case a dependent-causes while on treatment strategy, implemented with a competing risks approach, might be appropriate.

**Subjective clinical progression.** Oncology clinical trials are increasing requiring follow-up until documented radiological progression as clinical progression has been criticized as an ill-defined, potentially subjective, and potentially biased assessment. Many trials offer the option for participants meeting pre-specified criteria who are in the opinion of the investigator continuing to derive benefit to remain on treatment after the initial documentation of radiographical progression. In such cases where it is broadly feasible to followed participants until documented progression a treatment policy strategy is appropriate. However a composite strategy might be considered in cases where clinical progression prior to radiological progression is expected to occur frequently and it is not practical to continue to follow participants, particularly where clinical progression is highly predictive of radiological progression or death. In some cases, potential for bias due to subjective nature of clinical progression might need to be balanced against bias due to systemic inability to follow patients resulting in informative censoring.

**Central vs. local review.** In oncology trials, it is not uncommon to have a retrospective central radiological review in addition to a local review by the investigator. One of these reviews, generally the central review, is considered primary. There is often a discrepancy between central and local review results. When central review is retrospective, patient management decisions including decisions to discontinue tumor assessments are based on local review. When a patient progresses per local review, further imaging is not available to the retrospective central review, resulting in PFS being censored if the central review finds no progression. But when central review finds progression before the local review, both reviews are complete. Fleischer (2010) discussed the issue and hypothesized a model under which this difference introduces bias into the results. Zhang (2013), however, evaluated a meta-analyses of studies and concluded that retrospective central review introduces variation but not bias. We would note that both local and central review are addressing the same scientific question. The introduction of real-time central review, in which the central review decision is conducted rapidly and the results are considered by the investigator in making treatment decisions, can alleviate the issue, although the one-way transmission of information that occurs in the process can result in the central review results informing or replacing local review. In any event the secondary review should be considered as sensitivity analysis.

**Incorrect medication.** Receipt of the wrong study medication in a trial is generally a rare, isolated event. As such, even though patients with incorrect treatment are often discontinued from most study assessments, and resulting censoring may well be informative, these events are not likely to be pervasive or systematic enough to affect the overall interpretation of the results or choice of estimand. Accordingly, a treatment policy strategy can generally be used. In the unlikely event that incorrect medication is a pervasive issue in a study, a hypothetical strategy could be used to assess what would have happened if incorrect medication had not been given.
Initiation of new anti-cancer therapy. When the subsequent therapy given in a trial is of a sort that would be given in normal clinical practice if the product was approved, and if patients can generally be followed beyond subsequent therapy to the event of interest, a treatment policy strategy can be used.

However, an increasing problem in contemporary trials is a situation where the trial conditions induce behavior that would not otherwise exist in the clinic. Examples discussed in Manitz et al (2022) include:

- In blinded trials, when there are multiple experimental treatments being studied in the same class, patients assigned to the control arm may later enter another trial evaluating an experimental treatment in the same class.
- In open-label trials, patients not assigned to the desired treatment may immediately withdraw without receiving study treatment, or otherwise withdraw earlier than would be the case in regular clinical practice. For example, in Checkmate 037, 23% of patients randomised to the control arm did not receive assigned treatment compared with 1% randomised to active therapy.

Both of these situations would not normally occur in the clinic. Patients who have progressed on a therapy would not normally take a therapy in the same class as subsequent therapy. And patients who begin a treatment regimen for a serious cancer would not normally immediately switch without cause. These situations are therefore artifacts of trial conditions, based on the fact that in a trial, patients are randomly assigned a treatment which may not be the one they were hoping to receive, but nonetheless have the ability to “vote with their feet” if they do not like the assignment. They thus violate an implicit assumption of a treatment policy strategy, that the treatment pattern observed in the trial predicts the treatment pattern that will occur in the clinic in the event of approval.

Where the context of a clinical trial results in conditions and induces patient choices not likely to be repeatable in real-world post-approval clinical practice, a counterfactual hypothetical strategy may better reflect the scientific question feasibly addressable by such a trial. Such a hypothetical strategy might be considered, implemented with causal inference methods such as rank-preserving structural failure time (RPSFT) to estimate the outcome had patients not crossed over from control to active treatment; the 2-stage method to estimate the outcome had patients not crossed over at a specific disease-related time point such as progression; and inverse probability censoring weighting (IPCW) to estimate the outcome in the absence of new therapy.

These methods have been subject to criticism because of the strength of the assumptions required, particularly the assumption of no unmeasured confounding. While the reliability of the answer provided by hypothetical strategies may be criticized and indeed cannot be assured, they have the advantage of addressing a clinically relevant question. See Manitz (2022) for a more detailed discussion.
As ICH E9 R1 notes, “usually an iterative process will be necessary to reach an estimand that is of clinical relevance for decision making, and for which a reliable estimate can be made.” Subsequent therapy may represent an example of a conflict between clinical relevance and reliable estimation, requiring care in selecting a scientific question and intercurrent event strategy representing a reasonable balance between the two.

**Surgery and stem cell transplant.** Surgery may occur during oncology trials for multiple reasons. For example, surgery is a planned procedure in neoadjuvant trials, or as palliative or curative treatment in later stage disease. For an EFS endpoint, surgery could be a component of a composite strategy. Alternatively, if unplanned surgery is potentially curative, a treatment policy approach may be considered, following the patient beyond surgery until the event of interest. In the case of palliative surgery, this could be handled similarly to initiation of anti-cancer therapy described above. Similar considerations apply to stem cell transplant as an intervention. It may not always be possible to follow patients beyond surgery. Censoring at the time of surgery would constitute a hypothetical strategy, and would have issues similar to censoring for other subsequent therapy.

In addition to these typical intercurrent events which occur in oncology clinical trials, there might be intercurrent events observed during the course of a trial which could not be foreseen at the design stage and which cannot be controlled by study procedures, like the occurrence of a pandemic. Such cases require careful re-assessment of the estimand definition.

### 6. Practical consequences for study design and data collection

In the estimands framework, data collection is as important as design, as the estimand definition(s) in a trial determine how long rigorous data collection is required. Additionally the chosen estimator will generally require data on intercurrent events. This may require augmentation or even redesign of standard data collection systems and procedures.

Loss to follow-up is not itself an intercurrent event, but is often caused by intercurrent events. For this reason, special attention is needed in the study design to forecast the likely potential intercurrent events that will result in loss to follow-up. The CRF should capture relevant data documenting loss to follow-up, including identifying intercurrent events and their likely relationships to treatment effect that might be associated with loss to follow-up. Data collection should include whether the assessment occurred, reason for no assessment, and a causality (relatedness to study treatment). This causality assessment could be similar to the one commonly used to determine treatment relatedness of AEs. The list of reasons for no assessment should not be limited to predefined intercurrent events. The underlying reason should be captured in sufficient granularity to identify intercurrent events not being considered relevant at the design stage. This requirement is also applicable to documenting withdrawal from study treatment or change in therapy.

ICH E9 (R1) states: “A prospective plan to collect informative reasons for why data intended for collection are missing may help to distinguish the occurrence of intercurrent events from
missing data. This in turn may improve the analysis and may also lead to a more appropriate choice of sensitivity analysis. Implementing this prospective plan, documenting the existence and reasons for missed assessments, together with a causality assessment (relatedness to treatment), represents a significant change from current practice. However, collecting this additional data to support the needed sensitivity analyses is important in the context of the estimand framework. This recommendation may be particularly relevant to time-to-event estimands that are the basis of label claims.

Alternative approaches to data collection might be considered to increase follow-up beyond certain types of intercurrent events. For example, remote visits and digital devices can collect physical function and patient diary data to replace frequent clinic visits to complete questionnaires. Such automated and at-home data collection avoids additional burden for patients and minimizes the risk to stop follow-up prematurely. Improvements in technology have already aided data collection outside clinic visits and are likely to continue to improve.

7. Addressing when patients cannot be followed

Traditionally, oncology clinical trials ended clinic visits at events like treatment discontinuation or progression, with only long-term follow-up (typically telephone contacts) afterwards. In addition to the estimand framework, an evolving interest in understanding the impact of treatment on how patients feel and function as well as an evolving therapeutic landscape often requires additional data to understand whether or not the natural history of the disease is worsened by the study drug and to demonstrate if there can be benefit beyond initial disease progression. These concerns have resulted in an increased interest in following patients past intercurrent events. However, it is recognized that there are often practical limitations. Additionally, it is recognized that required follow-up within a clinical trial may focus on the primary or key secondary endpoints and in an effort to reduce burden on trial participants limited considerations may be given to follow-up for other supportive endpoints. For example, in many oncology trials follow-up post progression may be substantially reduced to focus on remote assessments to collect survival and subsequent treatment. In such cases continued follow-up on patient reported outcomes may be halted limiting the ability to understand the overall impacts of study treatment. Alternative to clinical visits as electronic diaries and home visits are increasingly being considered to allow for continued data collection.

The ICH E9(R1) guidance provides a useful framework for addressing cases where it may not be practical or feasible to continue follow-up. It provides multiple strategies and options for addressing intercurrent events, both options that require following patients beyond them, and options that do not. These options may provide a navigable path to defining a scientific question of interest that can be addressed in the context of the disease setting.

An additional practical issue involves subordination of follow-up for lower-priority estimands to the needs of higher-priority ones. It is not uncommon in oncology clinical trials for data collection considerations based on primary variable to affect data collection for secondary
variables. For example, if the primary variable is PFS, clinic visits, even when extending past subsequent therapy, will typically end at documented radiological progression. Ending clinic assessments at progression will occlude data collection for secondary estimands requiring in-clinic assessments, for example time to forced lung capacity deterioration and time to symptom improvement or deterioration.

Where assessment ends at an event based on the needs of a different, higher-priority estimand, and the ending event may be informative with respect to the event of interest, we recommend that the study design team explicitly acknowledge the event resulting in loss to further clinic visits as an intercurrent event, and devise an appropriate strategy that is both scientifically reasonable and feasible in the context. An explicit hypothetical strategy would be preferable to leaving the effect of the assessment-ending event unacknowledged. But other strategies acknowledging the impact of the event on interpretation may be more appropriate, particularly where the relevant event is highly informative. It might, for example, be appropriate to consider a composite strategy and assess time to the earlier of symptom deterioration or progression, whichever occurs first. In some cases, the lowered priority of the estimand may reflect the fact that what happens after the assessment-ending event is not of sufficient interest to be worth study in the context of the particular clinical trial, and a while on treatment (i.e. while prior to occluding event) strategy may be appropriate.

Compromises are inevitable in applied clinical research. It is important for the design team to first identify the clinical question of interest and understand the optimum strategy to address that question. Once that is done, the team should proceed to investigate whether the clinical conditions, estimand priority within the study, and other factors render the strategy feasible in the context. If a compromise is necessary, the team should be conscious of both what is desired and what is possible. It should understand the limitations compromise impose on the ability of the study to address the original question, including understanding how changing the strategy changes the research question, and how design features can implicitly change the strategy.

The Estimands Guidance notes that “usually an iterative process will be necessary to reach an estimand that is of clinical relevance for decision making, and for which a reliable estimate can be made.” Figure 1 illustrates a proposal for an iterative approach to construct an estimand that is both of clinical relevance for decision making and operationally feasible at the design stage. This iterative approach is similar to Deming’s Plan-Do-Study-Act cycle.28
Figure 1: Illustration of an iterative approach to construct an estimand that is of clinical relevance for decision making and operationally feasible.

8. **Sensitivity analyses**

ICH E9 (R1) defines sensitivity analyses as “a series of analyses conducted with the intent to explore the robustness of inferences from the main estimator to deviations from its underlying modeling assumptions and limitations in the data.” The purpose of a sensitivity analysis is to check the assumptions underlying the estimand. This concept of a sensitivity analysis represents a change in meaning from common past use. In the past, various standard statistical assumption checks, like checking for non-proportional hazards or informative censoring, might have been classified as part of the primary analysis rather than under the rubric of sensitivity analyses. The estimands framework classifies these checks as sensitivity analyses. Similarly, in the past a variety of different analyses involving e.g. alternative censoring rules or definitions of the event of interest might have been classified as sensitivity analyses. Under the estimands guidance, analyses that evaluate a different event of interest or a different strategy for addressing intercurrent events, such as various alternative ways of defining progression traditionally performed in analyses of PFS, are supplemental analyses, not sensitivity analyses.

Every strategy is based on assumptions and requires conditions to be valid. The purpose of sensitivity analyses is to check, to the extent feasible, whether the applicable assumptions
required for the primary estimand are reasonable under the circumstances. The discussion here focuses on sensitivity analyses, as defined in ICE E9 addendum, related to censoring. Sensitivity analyses are a particular problem in a survival context.

A key element of sensitivity analyses in a survival context is to check the appropriateness of non-informative censoring assumptions, and hence whether presumed “missing data” can in fact legitimately be so characterized. Events that occur after censoring are not always documented and it is not possible to know whether censoring is informative. Similarly, assumptions needed for a hypothetical analysis are often unverifiable. Nonetheless standard censoring checks and sensitivity analyses are possible, and commonly employed. Examples include:

- Assessing the distribution of censoring to see if it occurs evenly between the arms.
- Analyses with and without censoring of events that are observed following more than one missed assessment. While this analysis was traditionally performed regardless of reason for missed assessment, it might be appropriate to perform an additional analysis focusing on assessments missed due to possible intercurrent events (per CRF data collected as recommended in Section 6).
- Checks for key traditional statistical assumptions such as proportional hazards
- Interval censored analyses (potentially a different estimand and hence a supplementary analysis, but also can be used to check the appropriateness of right-censoring).

As discussed above, a common issue in survival trials is the use of treatment policy strategies where systematically following patients beyond the intercurrent event of concern may not be feasible. We recommend sensitivity analyses to address this assumption. This could be done with simple descriptive methods. For example, for PFS studies with a treatment policy strategy for events like end of treatment, clinical progression, or change of therapy without radiological progression, we recommend identifying what proportion of patients had these events prior to progression, and what proportion did not receive further tumor assessments beyond these events. Large differences in the number of events between arms could lead to concern. Sensitivity analyses could assess different assumptions for patients who were lost to follow-up due to other, e.g. unambiguously terminal events. For example, patients who died shortly after change of therapy could be assumed to have received full follow-up regardless of the change.

Supplementary analyses, by contrast, target a different estimand and provide additional insights into the treatment effect. For example, these could include exploration of the components of a composite endpoint and compare results from local and central tumour assessments, or various alternative definitions of progression (e.g. including clinical progression).
9. **Noumenal and phenomenal censoring**

Censoring has been characterized in multiple ways. It can be described as:

- A natural property of the underlying clinical trial context in which patients end follow-up
- A mathematical estimation technique for handling missing data in survival analysis, a component of the estimator,
- A component of the estimand.

When applying the estimands framework, there is a distinction between censoring as part of the estimator and censoring part of the estimand. New terminology may be helpful in distinguishing these concepts.

Censoring should be considered part of the estimand when it occurs systematically for a common event. This occurs when all patients are censored for the event as a matter of design per the visit schedule, withdrawal criteria, or analysis specification table that defines censoring rules. This type of systematic censoring can also occur in trial practice even if not originally foreseen. Censoring should be considered part of the estimator when it occurs for isolated patients and comparatively rare events. Systematic censoring affects the set of possible strategies that can be applied in a particular trial, while isolated censoring does not do this, even though it may sometimes cast doubt on the validity of the results.

To make this distinction explicit the authors propose terms borrowed from Kant, labelling systematic censoring for a common or pervasive event as *noumenal* censoring, and labelling censoring for isolated and non-systematic events as *phenomenal* censoring. In Kant’s terminology, a *noumenon* is an unobservable thing-in-itself; while a *phenomenon* is what appears to the senses. *Phenomena* in this framework are potentially affected by the act and apparatus of observation; *noumena* are not. We think that this terminology may be helpful because it is analogous to the idea of an estimand inferring to and describing a potentially unobservable, perhaps hypothetical population, with the estimator always calculated from the data observed in the trial. These terms help convey the idea that systematic, pervasive censoring can be thought of as part of the estimand itself (the *noumenon*), while isolated censoring, even if informative, can be thought of as part of the estimator (the *phenomenon*).

The distinction between noumenal and phenomenal is not quite the same as the distinction between non-informative and informative. Censoring for the pre-planned end of the trial is perhaps a paradigmatic example of uninformative censoring. We nonetheless classify this type of censoring as noumenal because it is systematic and pervasive. Similarly, isolated censoring events, even if highly informative, would be classified as phenomenal censoring, because they do not change the set of possible estimands or the interpretation of the data. An example would be occasional incorrect treatment (not as randomized) that results in the patient being withdrawn. This is a distinct intercurrent event; resulting censoring is informative; and yet if not pervasive it does not affect the overall interpretation of the results, so the censoring involved would be phenomenal.
These new terms may be helpful in assessing whether a treatment policy strategy is appropriate in the situation. In general, the treatment policy strategy, regardless of intercurrent events, has often become a de facto default standard and is often recommended for randomised pivotal clinical trials. However, as discussed above, a treatment policy strategy for a particular intercurrent event requires follow-up through and beyond the intercurrent event and to the event of interest. Generally speaking, where follow-up to the event of interest is infeasible, a treatment policy strategy is problematic, and inability to follow beyond the applicable event should be taken into account in constructing the applicable estimand. Since censoring resulting from a systematic inability to follow beyond an event represents noumenal censoring, we note that the presence of noumenal censoring generally means alternatives to the treatment policy strategy should be considered. It may be better interpreted as a simple hypothetical rather than a treatment policy strategy as noumenal censoring changes the estimand. This is why we argue that noumenal censoring should be considered part of the estimand, while phenomenal censoring remains part of the estimator.

When noumenal censoring is expected, alternative strategies include the composite strategy, the while-on-treatment strategy, and hypothetical strategy. In addition, a successful principal stratum strategy renders censoring for the applicable intercurrent event phenomenal, as it can be interpreted as representing isolated cases where the principal stratum model did not hold.

10. Discussion and conclusions

In the past, the assumption of non-informative censoring has rarely been challenged by regulatory authorities. Except in special cases like the pre-planned maturation of the trial, censoring is often informative. Its traditional widespread use to estimate in the presence of intercurrent events has often ignored the potential to bias results. The estimand framework addresses this issue because the concept of intercurrent events closely resembles the survival analysis concept of informative censoring. Hence the framework’s strategies for identifying and addressing intercurrent events provide methods for handling situations where the non-informative censoring assumption does not apply.

Different approaches are necessary to address different clinical questions of interest. As discussed in Section 6 the approach to study design should begin with the clinical question of greatest interest, but may require an iterative process to ensure alignment of the clinical question of interests with the estimand strategy, and to ensure that both can be validly supported by the data that can be feasibly collected in the planned trial.

The terms noumenal and phenomenal censoring are useful to distinguish between censoring that changes the interpretation of the results and hence can be thought of as part of the estimand, and censoring that, although potentially informative, does not affect the interpretation of the results and can be thought of as part of the estimator.

While the noumenal vs. phenomenal distinction affects the interpretation of the data and the research questions that can be addressed by it, the non-informativity assumption underlying
the use of censoring affects the validity of the results for the chosen strategy. Censoring for intercurrent events is often informative, and the validity of the non-informativity assumption should be checked.

In general, the treatment policy strategy, regardless of intercurrent events, has become the default standard and is recommended for randomised pivotal clinical trials where data can be consistently and systematically collected until the event of interest or study termination. From an estimands perspective, the treatment policy strategy reflects the entire treatment regimen, including subsequent therapies. Other strategies such as composite, hypothetical, or while-on-treatment strategy might be considered when the question of interest differs or where the assumptions underlying the treatment policy are not met. A treatment policy strategy will not insulate the trial from confounding, and events such as subsequent therapy that are inconsistent with real-world practice might better be handled as confounding intercurrent events than treated as censorable. It is important to recognize that changing the strategy and the handling of intercurrent events changes the estimand and its interpretation.

When the treatment policy strategy is not used, censoring is often replaced by another implementation mechanism. For example, composite strategies will handle events as a component of the event of interest; while-on-treatment strategies may handle them as a competing risk event; hypothetical strategies may implement a causal inference model; and so on.

Our paper emphasizes the problem of implicit noumenal censoring, which is an important problem in clinical trial practice. A study may be described as using a treatment policy strategy, but in fact on closer examination the study design shows that the protocol or patient-management practice results in patients being systematically withdrawn for some relevant intercurrent event, often radiological or clinical progression. We stress that a treatment policy strategy does not consist of merely applying the label and defining censoring rules in the analysis that does not explicitly censor for subsequent therapy or other intercurrent events. Rather it requires that patients actually can be and are followed, systematically, to the event of interest. Implicit noumenal censoring arises when there is no explicit censoring but patients are in fact generally censored at a particular intercurrent event. This type of censoring induces an implicit hypothetical strategy. By the time the trial is over it is far too late to be determining what strategy was used. Trialists should reduce the occurrence of implicit noumenal censoring in registrational clinical trials by making it explicit. Events which terminate assessments or trigger withdrawal criteria, such as progression in many studies, are particular candidates. When the study design systematically stops assessments at a particular event, we recommend explicitly identifying that event and determining an appropriate strategy to be used for that event. Attempting to apply catch-all censoring at end of assessments in situations where patients are being pervasively and systematically removed from follow-up by design or widespread practice, without investigating effect on research question or interpretation, is incompatible with the estimands framework.
It is better to ensure alignment of study design with goals and the requirements of the corresponding estimand up-front than to check for problems resulting from non-alignment later. Alignment requires a dialogue among clinicians, statisticians, and trial managers on the design team to clarify the research question and discuss how proposed withdrawal criteria, visit schedules, and anticipated patient and investigator behaviors could affect the interpretation of the study and limit the possible estimands that could be applied. This conversation should take place early, before study design proceeds very far.

The purpose of a clinical trial is often to predict real-world clinical practice, especially for a registrational trial. Certain elements of a clinical trial, such as randomization, are not reflective of real-world practice, and may induce patient behaviour not replicable in the real world. Where this occurs, a treatment policy strategy might not be the most appropriate. A counterfactual hypothetical strategy, which asks what would have happened if the non-real-world behaviour had not occurred, might be more relevant to real-world practice and hence might sometimes be preferable, despite problems with establishing the reliability of causal inference methods.

Rigorous data collection and trial monitoring are the key to addressing and distinguishing missing data and intercurrent events. The estimands framework depends on good data collection, including data about missed assessments. Events that are not collected cannot be managed or addressed. It is therefore critical to obtain the reasons for and dates of withdrawals and losses to follow-up, and to obtain data permitting assessment of the existence and dates of underlying intercurrent events.

While not described within the estimand guidance, one could consider framing the concept of censoring in the broader context of occluding events, with occlusion representing any loss to further follow-up and/or removal of further collected data from analysis. Occlusion constitutes a broader concept than the estimand guidance’s “intercurrent event” and “terminal event” and one appropriate to addressing a broader set of time-to-event methods in an estimands context.

The estimands framework is not simply new language to describe conventional practices. It requires a rethinking of study conception, design, planning, execution, and analysis. We anticipate it will have impact on the data collection and interpretation of most if not all oncology studies with time-to-event analyses.

11. Data availability statement
No new data is presented in this manuscript. This manuscript is based solely on previously published results.

12. References
1. International Conference on Harmonization. ICH E9(R) - Addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. (2019)

2. Akacha et al. Estimands in clinical trials – broadening the perspective. Stat in Med 2017; 36:5-19.

3. Ratich B et al. Choosing estimands in clinical trials: Putting ICH E9(R1) into practice. Therapeutic Innovation & Regulatory Science 2019; 20:1-18.

4. Malinckrodt et al. Estimands, Estimators, and Sensitivity Analyses in Clinical Trials. Chapman & Hall/CRC 2020.

5. Rufibach K. Treatment effect quantification for time-to-event endpoints—Estimands, analysis strategies, and beyond. Pharm Statist 2019; 18:145-165.

6. Larkin J, et al. Overall survival in patients with advanced melanoma who received Nivolumab versus Investigator's choice chemotherapy in CheckMate 037: a randomized, controlled, open-label phase III trial. J Clin Oncol. 2018; 36:383-390. https://doi.org/10.1200/JCO.2016.71.8023

7. Manitz J et al. Estimands in clinical trials with treatment switching in oncology. Pharm Statist 2022; 21:150-162.

8. Arnold M et al. Progress in cancer survival, mortality, and incidence in seven high-income countries 1995–2014 (ICBP SURVMARK-2): a population-based study. Lancet Oncol 2019; 20: 1493-505.

9. United States Food and Drug Administration. Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics: Guidance for Industry; 2018.

10. European Medicines Agency. Guideline on the evaluation of anticancer medicinal products in man. Methodological consideration for using progression-free survival (PFS) or disease-free survival (DFS) in confirmatory trials. Appendix 1; 2013.

11. Cheson et al. Revised recommendations of the International Working Group for Diagnosis, Standardization of Response Criteria, Treatment Outcomes, and Reporting Standards for Therapeutic Trials in Acute Myeloid Leukemia. J Clin Oncol 2003; 21:4642-9.

12. United States Food and Drug Administration. Acute Myeloid Leukemia: Developing Drugs and Biological Products for Treatment. Draft Guidance for Industry; 2020.
13. Carroll KJ. Analysis of progression-free survival in oncology trials: Some common statistical issues. *Stats Med* 2007; 6:99-113.

14. Lipkovich I, Ratitch B, and Mallinckrodt CH. Causal inference and estimands in clinical trials. *Stat Biopharm Res.* 2020; 12:54–67. https://doi.org/10.1080/19466315.2019.1697739

15. Fleming, T, Rothmann, M and Lu H. Issues in using progression-free survival when evaluating oncology products. *J Clin Oncol* 2009; 27:2874-2880.

16. Austin PC, Fine JP. Practical recommendations for reporting Fine-Gray model analyses for competing risk data. *Stat Med* 2017; 36:4391-4400.

17. Unkel, S. 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 Statist* 2019; 18:166–183.

18. Fine J, Gray R. A proportional hazards model for the subdistribution of a competing risk. *J Am Stat Assoc* 1999; 94:496-509

19. Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics* 2002; 58:21-29.

20. Bornkamp et al. Principal stratum strategy: Potential role in drug development. *Pharm Statist* 2021; 21:1-15.

21. Smith, M et al. Addition of radium-223 to abiraterone acetate and prednisone or prednisolone in patients with castration-resistant prostate cancer and bone metastases (ERA 223): a randomised, double-blind, placebo-controlled, phase 3 trial. *Lancet Oncol* 2019; 20:408-419.

22. Dodd L et al. Blinded independent central review of progression-free survival in phase III clinical trials: Important design element or unnecessary expense? *J Clin Oncol* 2008; 26:3791-3796.

23. Fleischer F, Gaschler-Markefski B, Bluhmki E. How is retrospective independent review influenced by investigator-introduced informative censoring: a quantitative approach. *Stat Med* 2011; 30:3373-86.

24. Zhang, J, et al. Assessment of audit methodologies for bias evaluation of tumor progression in oncology clinical trials. *Clin Cancer Res* 2013; 19:2637–45
25. Robins JM, Finkelstein DM. Correcting for noncompliance and dependent censoring in an AIDS clinical trial with Inverse Probability of Censoring Weighted (IPCW) log-rank tests. *Biometrics* 2000; 56:779-788.

26. Latimer NR, et al. Adjusting for treatment switching in randomised controlled trials – a simulation study and a simplified two-stage method. *Stat Methods Med Res* 2017; 26:724-751.

27. Sun S, Weber HJ, Butler E, Rufibach K, Roychoudhury S. Estimands in hematologic oncology. *Pharm Statist* 2021; 10.1002/pst.2108.

28. Deming, WE. *Out of the Crisis*. MIT Press; 1986.

29. Kant I. *Kritik der reinen Vernunft*; 1781.

30. Stang, Nicholas F., Kant’s transcendental idealism. *Stanford Encyclopedia of Philosophy (Spring 2021 Edition)*, Edward N. Zalta (ed.) [https://plato.stanford.edu/entries/kant-transcendental-idealism/](https://plato.stanford.edu/entries/kant-transcendental-idealism/)

31. Degtyarev E et al. Assessing the Impact of COVID-19 on the Clinical Trial Objective and analysis of oncology clinical trials - application of the estimand framework. *Stat Biopharm Res* 2020; 12:427-437.