Received: 3 August 2021 | Revised: 2 March 2022 | Accepted: 14 March 2022

DOI: 10.1002/sim.9398

TUTORIAL IN BIOSTATISTICS

Statistics in Medicine WILEY

Translating questions to estimands in randomized clinical trials with intercurrent events

Mats J. Stensrud[1] | **Oliver Dukes**[2,3]

[1]Department of Mathematics, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

[2]Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[3]Department of Applied Mathematics, Statistics and Computer Science, Ghent University, Ghent, Belgium

Correspondence
Mats J. Stensrud, Department of Mathematics, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland.
Email: mats.stensrud@epfl.ch

Abstract

Intercurrent (post-treatment) events occur frequently in randomized trials, and investigators often express interest in treatment effects that suitably take account of these events. Contrasts that naively condition on intercurrent events do not have a straight-forward causal interpretation, and the practical relevance of other commonly used approaches is debated. In this work, we discuss how to formulate and choose an estimand, beyond the marginal intention-to-treat effect, from the point of view of a decision maker and drug developer. In particular, we argue that careful articulation of a practically useful research question should either reflect decision making at this point in time or future drug development. Indeed, a substantially interesting estimand is simply a formalization of the (plain English) description of a research question. A common feature of estimands that are practically useful is that they correspond to possibly hypothetical but well-defined interventions in identifiable (sub)populations. To illustrate our points, we consider five examples that were recently used to motivate consideration of principal stratum estimands in clinical trials. In all of these examples, we propose alternative causal estimands, such as conditional effects, sequential regime effects, and separable effects, that correspond to explicit research questions of substantial interest.

KEYWORDS

causal inference, estimands, identification, intercurrent events, principal stratification

1 | INTRODUCTION

Defining, interpreting, and identifying causal effects in the presence of an intercurrent (post-treatment) event is not straightforward, even in a randomized controlled trial (RCT). It is well-known that randomization of a baseline treatment does not ensure identification of estimands that are (implicitly or explicitly) defined conditional on a post-treatment variable. Moreover, a naive contrast of outcomes conditional on the post-treatment variable does not have a clear causal interpretation.

In this article, we discuss key issues concerning the formulation and choice of a causal estimand in settings with intercurrent events. Our first message is that an explicit research question should always precede the choice of estimand: it is the question that motivates the estimand, not the estimand that motivates the question. While this point may seem obvious and has been pointed out several times in the causal inference literature, numerous studies fail to give such a

motivation. As a result, the interpretation—and practical relevance—of these analyses is ambiguous. Furthermore, an implication of this message is that the choice of estimand critically depends on the context; an estimand that is relevant in one study may be irrelevant in another.

To illustrate the importance of carefully defining a causal estimand, we revisit five recent examples[1] inspired by the International Council of Harmonization (ICH) E9 (R1) addendum.[2] These recently published guidelines stress the importance of choosing a treatment effect (estimand) in a clinical trial that is well aligned with the clinical question of interest, with specific attention given to the handling of intercurrent events. As will become clear in Sections 3 to 5, the intercurrent events in these examples do not prevent a classical intention-to-treat (ITT) effect from being identified. However, decisions makers and drug developers will often be interested in additional questions that translate to different estimands. We argue that the causal estimands of interest in these secondary analyses should be selected on a case-by-case basis.

Furthermore, we clarify that the translation of a research question into a formal estimand is separate from the question of its identification, that is, whether and how it can be expressed as a functional of distributions of observables. Establishing plausible conditions for identification can be considered a distinct—but important—task, which should be carefully conducted *after* choosing the causal estimand. The final (distinct) task is estimation, for which one may choose from more or less parametric or robust approaches. In other words, we find it helpful to distinguish between the following three tasks of data analyses: (i) translation of the research question into a formal causal estimand, (ii) assessing conditions for identification of the causal estimand, and (iii) estimation of the causal estimand from observed data. In this article, we will focus on task (i), but also briefly consider tasks (ii) and (iii).

We have structured our arguments as follows. In Section 2, we briefly introduce the intercurrent event settings in the five clinical examples from Bornkamp et al.[1] In Section 3, we describe estimands that have been suggested for causal inference in clinical settings with intercurrent events.* In Section 4, we revisit the clinical examples and map the subject matter questions—as described in plain English by the investigators—to causal estimands. In Section 5, we review conditions that allow us to identify these estimands in an RCT, followed by a brief description on how to estimate them in Section 6. In Section 7, we give a discussion.

2 | CLINICAL EXAMPLES

Example 1 (Multiple sclerosis). Multiple sclerosis (MS) is a progressive neurological disease. Initially most patients have a phase with relapses followed by recoveries, and eventually the patients transform to a secondary phase with less frequent relapses. There is major interest in developing new treatments that delay or prevent disease progression. For example, the EXPAND study was a randomized clinical trial that assigned the drug siponimod vs placebo to patients in the secondary phase, where the primary estimand was the onset of confirmed disability progression.[3] Siponimod was shown to delay the onset of disability progression compared to placebo, and it was also shown to reduce the frequency of relapses. These primary results raised the question whether the treatment could affect disability progression outside of its effects on relapses.[1,4] Here, experiencing relapses is an intercurrent event.

Example 2 (Treatment effects in early responders). A biomarker is a variable that quantifies a biological state. Treatment effects on biomarkers, such as high sensitivity c-reactive protein,[5] can serve as early predictors of treatment effects on clinical outcomes. The ASA/EFSPI oncology estimand working group[1] further suggested that "biomarkers or early readouts can be useful to investigate whether an investigational medicine works as intended on a biological level." Thus, a motivation for studying biomarker responses in RCTs seems to be elucidation of causal mechanisms (on a biological level), which can indicate whether the drug acts as intended through particular causal pathways. Here, reaching a biomarker threshold is an intercurrent event.

Example 3 (Impact of exposure on overall survival). In drug trials, individuals who are given the same dose of a drug can still have different concentration of the active substance of this drug in the blood (serum). The concentration of the active substance can be important for the treatment effect. An RCT that assigned Trastuzumab for gastric cancer showed that those patients in the quartile with the lowest drug concentration had worse overall survival (OS) compared to the other quartiles.[6] Researchers at the Food and Drug Administration[7] subsequently raised the question "whether the lower OS is due to low drug concentration or to disease burden" (see also References 1 and 8). Here, the drug concentration is considered to be an intercurrent event.[1]

Example 4 (Antidrug antibodies for targeted oncology trials). Immunotherapies are increasingly used to treat several cancers. Yet, there is concern that some immunotherapies can trigger the production of antidrug antibodies (ADAs),

which in turn can reduce the treatment effect of the immunotherapies on, say, overall survival.[9] Indeed, "ADAs may be directed against immunogenic parts of the drug and may affect its efficacy or safety, or they may bind to regions of the protein which do not affect safety or efficacy, with little to no clinical effect."[1] Here, the production of ADAs is an intercurrent event.

Example 5 (Prostate cancer prevention). There is major interest in developing drugs that prevent development of diseases, such as cancer. In particular, finasteride was shown to reduce the rates of prostate cancer in an RCT.[10] While there were lower rates of cancer in the finasteride arm, those who developed cancer in the finasteride arm had, on average, more aggressive cancers than those assigned to placebo. This conditional association, however, does not necessarily have a causal interpretation—it could just be that the more aggressive cancers cannot be prevented by finasteride while mild ones can be prevented. Thus, after establishing that finasteride reduces the rate of prostate cancer, Bornkamp et al[1] suggested a secondary analysis to assess whether "the effect of finasteride on the severity of prostate cancer among those men who would be diagnosed with prostate cancer regardless of their treatment assignment." Here, being diagnosed with prostate cancer is an intercurrent event.

2.1 | Common characteristics of intercurrent events in Examples 1 to 5

In Examples 1 to 5, the intercurrent event does not render the outcome of interest ill-defined. That is, we can evaluate the effect of siponimod vs placebo on disability progression in the entire study population in Example 1, whether or not patients have relapses. Similarly, we can evaluate the effects of canakinumab on MACE (Example 2), Trastuzumab on gastric cancer (Example 3), immunotherapies on cancer progression (Example 4), and finasteride on prostate cancer incidence (Example 5), whether or not individuals experience the intercurrent events. Thus, we can study total effect estimands in all of the examples, but nevertheless there may be interest in endpoints that take into account the intercurrent events.

Furthermore, we do not know about any intervention that fixes the intermediate event in Examples 1 to 5. In particular, we do not have any current or future treatment that fixes MS patients to relapse or not relapse.

Thus, Examples 1 to 5 have two important characteristics: (i) we cannot conceive plausible interventions on the intercurrent events, and (ii) the occurrence of the intercurrent event does not render the outcome of interest ill-defined. Other intercurrent events do not necessarily share these characteristics. For example, Michiels et al[11] studied treatment effects in clinical studies where patients sometimes received rescue medications in response to worsening of a disease. Here, the use of rescue medication was considered to be an intercurrent event, and it is easy to conceive an intervention to give or reject rescue treatment.[†] Furthermore, some outcomes are often considered to be ill-defined in the presence of an intercurrent event. For example, many researchers consider quality of life to be ill-defined after death, where death can be considered to be an intercurrent event. Whether or not events render the outcome ill-defined for certain individuals and are interveneable, has implications for the practical relevance and plausibility of estimands, as we discuss in more detail in Section 3 and Appendix B.

3 | NOTATION AND ESTIMANDS FOR SETTINGS WITH INTERCURRENT EVENTS

3.1 | Average total effects

The primary estimand in most randomized trials is the (average) total effect of treatment assignment. This is called the *treatment policy strategy* estimand in the ICH addendum,[2] and it coincides with the ITT effect in case of nonadherence. For simplicity, we will consider average effects, often on the additive scale, in the remainder of this article. Such averages of individual level effects are often the most relevant for decision making at this point in time and have a clear causal interpretation. However, our conceptual points are also valid for other causal contrasts.

To formally define this estimand, consider a study where individuals are randomly assigned to a binary treatment $A \in \{0, 1\}$ at baseline; while, strictly, it is the assignment that is randomized, we will simply speak of the randomized treatment A in the following. Let Y be the outcome of interest measured at a fixed time $t > 0$, and let S be an indicator of an intercurrent event that occurs at a time b such that $b < t$. Thus, for the ease of exposition and to align with recent considerations of intercurrent events,[1,11,12] we have simplified Y and S to be time-fixed variables, and we suppose that the

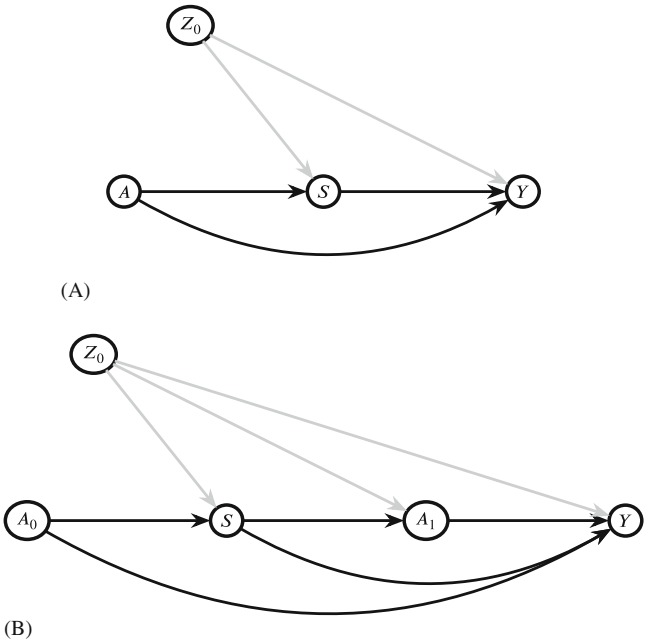

FIGURE 1 Causal DAGs that describe a randomized trial where baseline treatment A is randomly assigned (A), and a sequential randomized trials where A_0 and A_1 are randomly assigned, where the assignment of A_1 depends on the intercurrent event S. The intercurrent event S and the outcome of interested Y may be affected by common causes Z_0

event of interest occurs after the intercurrent event time b. However, most of our arguments can be extended to settings where the intercurrent event and the outcome are time-varying, for which theoretical results exist.[13-19] A consequence of our time fixed set-up is that the outcome of interest must be defined after the intercurrent event.

A causal DAG that is consistent with our setting is show in Figure 1A, where randomization of A ensures there are no common causes of A and Y. Let superscripts denote potential outcomes, such that Y^a and S^a denote the potential outcome of interest and the post-treatment event, respectively, had an individual, possibly contrary to the fact, been assigned $A = a$, where $a \in \{0, 1\}$. The average total effect of the treatment A on the outcome Y is a contrast

$$\mathbb{E}(Y^{a=1}) \text{ vs } \mathbb{E}(Y^{a=0}). \tag{1}$$

The average total effect (1) compares the average outcome in the population had everyone been treated ($a = 1$) vs not treated ($a = 0$); it ignores any intercurrent events. In all of the examples of Section 2, the average total effect of treatment assignment is well defined and meaningful. It describes the effect of assigning siponimod vs placebo on disability (Example 1), canakinumab vs placebo on MACE (Example 2), immunotherapies vs placebo on cancer progression (Example 3), Trastuzumab vs placebo on gastric cancer in patients receiving chemotherapy (Example 4), and finasteride vs placebo on prostate cancer incidence (Example 5), without considering any intercurrent events. The intercurrent events in these examples are such that the treatment/control and the outcome remain meaningful with or without the intercurrent event. In other situations this total effect will not necessarily be of interest, as discussed in the ICH addendum, such as "discontinuation of assigned treatment, use of an additional or alternative treatment, drop-out and terminal events such as death"[2(SectionA.1)] (See also References 19 and 20). Nonadherence, or changes in treatment such as rescue treatment, may imply a substantial modification of intended treatment which is why an ITT analysis is sometimes regarded as unsatisfactory; while competing events or drop-out may make the outcome impossible to occur or be measured.

Consider now a setting where treatments are *sequentially* given at multiple times, which motivates the study of sequential treatment effects.[21,22] For example, let A be an immunotherapy, and suppose that the doctor every week (sequentially) recommends whether a patient should initiate, continue or discontinue the therapy. To fix ideas, suppose that treatment can be given at baseline (time 0) and one subsequent point in time (time 1), and let A_0 and A_1 be treatment indicators at these times (Figure 1B). The causal effect comparing a sequential treatment strategy that fixes $A_0 = a_0$ and then $A_1 = a_1$ vs an alternative strategy that fixes $A_0 = a_0'$ and then $A_1 = a_1'$ is

$$\mathbb{E}(Y^{a_0,a_1}) \text{ vs } \mathbb{E}(Y^{a_0',a_1'}), \tag{2}$$

where $a_0, a_1, a_0', a_1' \in \{0,1\}$. Importantly, the sequential trial estimand is often of interest, even in clinical trials that assign a single treatment strategy at baseline. For example, formal (causal) definitions of per protocol estimands,[23] which arguably are useful in a range of practical settings such as pragmatic trials,[22,24] often require specification of sequential treatment regimes: the per protocol estimands evaluate the effect of taking treatment, as described in the protocol, at *every* point in time vs not taking treatment at *every* point in time. Alternatively, we might consider the effect of treatment received at baseline, if no one (or everyone) received treatment subsequently. The *hypothetical strategy estimand* referred to in the ICH addendum[2] can thus also be viewed as a sequential trial estimand.

Estimand (2) is a special case of a broader class of so-called dynamic sequential treatment regimes. These regimes are called dynamic, because the sequential treatment decisions can depend on each patient's previous treatment and other characteristics. In other words, the treatment decision at time k can be a function of (time-varying) covariates up until time k, and these covariates can include the intercurrent event status and treatment that have been received previously. For example, a decision of continuing a medical treatment at a time k may depend on the clinical history up until time k, previous treatments that were received and potential side-effects.

Let $g \in \mathcal{G}$ be a regime that fixes the treatment at two points in time: First A_0 is set to a_0 and subsequently A_1 is set to $a_1 = f_g(a_0, s)$ where $f_g(\cdot)$ is a deterministic function of the first treatment a_0 and the intercurrent event $s \in \{0,1\}$: that is, a regime g where the treatment at time 1 depends on whether the patient received the treatment at time 0 and the status of the intercurrent event S before time 1. For example, suppose g is defined such that only individuals who received treatment at time 0 and did not experience the intercurrent event will receive treatment at time 1.

We can define a contrast of two regimes $g, g' \in \mathcal{G}$,

$$\mathbb{E}(Y^{g}) \text{ vs } \mathbb{E}(Y^{g'}). \tag{3}$$

The total effect (1) and the sequential trial estimands (2) and (3) do not quantify the mechanisms by which the treatment affects the outcome Y. In particular, these estimands neither quantify effects conditional on nor mediated through the intercurrent events. On the other hand, these estimands are immediately relevant for designing practically feasible treatment regimes for the existing treatment A, which can be given at different time points. Clinicians and patients will be particularly interested in the optimal regime $g^* \in \mathcal{G}$ that leads to the most favorable expected clinical outcome, that is,

$$g^* \equiv \arg\max_{g \in \mathcal{G}} \mathbb{E}(Y^{g}). \tag{4}$$

The average total effects (1) and the sequential trial estimands (2) and (3) compare (counterfactual) outcomes under different treatment assignments in the entire study population. These estimands can easily be modified to average effects conditional on (functions of) observed pretreatment variables Z, such as

$$\mathbb{E}(Y^{g}|Z) \text{ vs } \mathbb{E}(Y^{g'} \mid Z),$$

and we discuss such conditional effects in more detail in Section 4. Alternatively, we can define conditional causal effects of the second treatment, given the first treatment and other events before time k. The causal effect of assigning $A_1 = a_1$ vs $A_1 = a_1'$ among those who received baseline treatment a_0 and had intercurrent event status s is

$$\mathbb{E}(Y^{a_0,a_1} \mid S^{a_0} = s) \text{ vs } \mathbb{E}(Y^{a_0,a_1'} \mid S^{a_0} = s). \tag{5}$$

The estimand (5) only quantifies effects of the second, but not the first, treatment assignment. Furthermore, this estimand is restricted to the subpopulation that had intercurrent event status s under treatment a_0.

3.2 | Principal stratum effects

The term principal stratification was coined by Frangakis and Rubin as follows:[25] "*Principal stratification with respect to a posttreatment variable is a cross-classification of subjects defined by the joint potential values of that posttreatment variable*

under each of the treatments being compared." This idea was introduced by Robins[21] when he considered counterfactual outcomes in individuals who would survive, regardless of treatment assignment.[‡] To fix ideas about principal stratum effects in a clinical trial setting, consider a randomized trial where a binary treatment $A \in \{0, 1\}$ is assigned at baseline. Let $Y \equiv Y(t)$ be the outcome of interest measured at a fixed time $t > 0$, and let $S \equiv S(b)$ be an indicator of a post-treatment event defined at a fixed time b, where $0 < b < t$. The additive principal stratum effect in the stratum defined by $S^{a=1} = s, = S^{a=0} = s'$ is

$$\mathbb{E}(Y^{a=1} - Y^{a=0} \mid S^{a=1} = s, S^{a=0} = s'). \tag{6}$$

Note that there are *joint potential outcomes* in the conditioning set of (6); this effect is defined in the subset of individuals characterized by the counterfactual intercurrent events $S^{a=1} = s$ and $S^{a=0} = s'$. In the example of a vaccine trial (suggested in the ICH addendum[2]), one might consider the treatment effect in those who would be infected regardless of vaccine assignment (say, $S^{a=1} = s, = S^{a=0} = s$). Because this subpopulation has the same intercurrent event regardless of treatment, the principal effect can be interpreted as both a direct effect of the vaccine (outside of infection) and a total effect.

The fact that (6) is defined with respect to joint potential outcomes has raised concern in the causal inference literature.[15,21,28-33] In general, these joint potential outcomes cannot be observed in the same individual, and therefore the principal stratum effects are defined in an unobserveable subpopulation that may not even exist, although it may sometimes be possible to obtain informative bounds on the size of the principle strata. Furthermore, identifying the (members of) principal strata are plausible in certain special settings, in which the principal stratum estimands correspond to questions of clinical interest.[34-36]

The ICH addendum[2] uses a broader definition of principal stratum effects, which encompasses estimands that condition on combinations (unions) of principal strata. In particular, the contrast

$$\mathbb{E}(Y^{a=1} - Y^{a=0} \mid S^{a=1} = s), \tag{7}$$

where the conditioning set is the union $(S^{a=1} = s, S^{a=0} = 1) \cup (S^{a=1} = s, S^{a=0} = 0)$ is, in addition to (6), included in the ICH definition of principal stratum estimands. The conditioning set in (7), that is, the union of principal strata, can be identified from a randomized trial. For example, in a vaccine trial we can look at the subpopulation in the vaccine arm who were infected. Yet, it does not necessarily express a direct effect.[§] In particular, (7) possibly quantifies effects both through and outside of the intercurrent event S unless it is assumed that the treatment A does not affect S.[15]

3.3 | Causal effects that reflect mechanisms

In the presence of intercurrent events, investigators often raise questions about the mechanisms by which the treatment affects the outcome of interest; that is, whether the treatment affects the outcome of interest through or outside of the intercurrent events. To clarify why such questions are of substantial interest, our experience is that investigators give stories about modified treatments that leverage certain mechanisms of the original treatment.[13,14,16,19,20,28,37,¶] The motivation seems to be that by understanding the causal mechanisms by which the current treatment affects the outcome of interest, we can motivate new, improved treatments in the future. Here we will define a class of mechanistic estimands called separable effects,[13-16,28,37] inspired by the seminal treatment decomposition idea from Robins and Richardson,[28] that are particularly helpful in this setting. To fix ideas about separable effects, we start with an example.

Example (Statins)**.** Statins are one of the most commonly prescribed drug classes in the world. They are successful because they reduce the risk of cardiovascular disease in a broad range of individuals with various clinical histories. The main mechanism by which statins reduces cardiovascular risk is lowering of low-density lipoprotein (LDL) cholesterol. However, a substantial fraction of patients who take statins experience muscle symptoms, which represent a big hurdle and can lead to treatment discontinuation.[41] More recently, new classes of drugs have been developed to specifically lower LDL levels like statins, but nevertheless differ from statins in the effects through other biological pathways. In particular, protein convertase subtilisin/kexin type 9 (PCSK9) inhibitors selectively reduce LDL levels, but PCSK9 inhibitors

do not exert effects through other biological pathways that is affected by statins (and can lead to muscle symptoms). Indeed, PCSK9 inhibitors have been successfully shown to reduce cholesterol levels in patients with muscle related statin intolerance.[42]

3.3.1 | Separable effects are effects of modified treatments

The separable effects are designed to target effects of (future) treatments, which selectively exert effects through certain (desirable) causal pathways similarly to the original treatment (eg, the LDL lowering induced by statins), but also selectively avoid other (undesirable) causal pathways (eg, non-selective pathways that can lead to muscle pain). In Section 4, we consider the potential role separable effects in the examples from Section 2, where new (improved) drugs remain to be discovered.

More abstractly, the separable effects are defined with respect to modified treatment components, A_Y and A_S, which are linked to the original treatment A in the following way: when A_Y and A_S are set to the same value a, the effects of giving this combination of modified treatments ($A_Y = A_S = a$), is the same as the effects of giving the original treatment $A = a$ for $a = 0, 1$ (see previous works[13-16,28,37] and Appendix A for a more formal derivation). This modified treatment assumption holds when the original treatment A can be decomposed into two components A_Y and A_S. However, this assumption can also hold even if A cannot be physically decomposed.[15] When the components A_Y and A_S are given individually, they exclusively target certain causal pathways (see Appendix A for details).

In practice, a study of separable effects should be motivated by a scientific story about modified treatments A_Y and A_S; the investigators should clarify why modified treatments are of scientific interest. One reason could be to develop improved treatments in the future. For example, let A be statin therapy and Y be cardiovascular risk. The combination $A_Y = 1$ and $A_S = 0$ can be conceived as a modified therapy (such as PCSK9 inhibitors) that, like statins, have a cholesterol lowering component $A_Y = 1$, but lacks effects on the intercurrent event such as muscle symptoms, $A_S = 0$.

More explicitly, we define the separable effect of the A_Y component as

$$\Pr(Y^{a_Y=1,a_S} = 1) \text{ vs } \Pr(Y^{a_Y=0,a_S} = 1), \quad a_S \in \{0, 1\}, \tag{8}$$

which quantifies the causal effect of the A_Y component on the risk of the outcome under an intervention that assigns $A_S = a_S$. Similarly,

$$\Pr(Y^{a_Y,a_S=1} = 1) \text{ vs } \Pr(Y^{a_Y,a_S=0} = 1), \quad a_Y \in \{0, 1\}, \tag{9}$$

quantifies the causal effect of the A_S component on the event of interest under an intervention that assigns $A_Y = a_Y$. The separable effect (8) quantifies the direct effect of the treatment on the outcome of interest, but this direct effect is different from the principal stratum estimand because it is defined in the entire study population. Similarly, (9) quantifies the indirect effect of the treatment on the outcome of interest through the intercurrent event. This decomposition illustrates that separable effects quantify mechanisms, and, unlike the sequential trial or principal stratification estimands, the separable indirect effect offers a coherent notion of an indirect effect.

In settings with intercurrent events, researchers often express interest in estimands in subgroups defined by the intercurrent event. This is particularly relevant when the outcome of interest only is well defined conditional on the status of the intercurrent event (eg, when there is truncation by death). Following Stensrud et al,[15] under the assumption that A_Y partial isolation (A1) holds, we can also define a conditional separable effect as

$$\mathbb{E}(Y^{a_Y=1,a_S} - Y^{a_Y=0,a_S} \mid S^{a_S} = s). \tag{10}$$

The conditional separable effect is the average causal effect of the (modified) treatment A_Y on Y when all individuals are assigned the other (modified) treatment $A_S = a_S$ in those individuals who have intercurrent event status s under $A_S = a_S$ (regardless of their value of A_Y).

3.4 | Relations between estimands

Like conventional mediation estimands,[13,28] such as natural (pure) direct and indirect effects,[43] the (marginal) separable effects quantify the mechanisms by which the treatment affects the outcome of interest.[13,14,16,28,37] Unlike the conventional mediation estimands, the separable effects do not require specification of any intervention on the intermediate event. This is important in our context, because it is not clear that such interventions on the intercurrent events exist in any of Examples 1 to 5. Furthermore, a feature of the separable effects is that they are defined with respect to modified treatments, and therefore they can be directly relevant in a drug development setting where the interest is tailoring new treatments to include/exclude certain pathways. However, as we return to in Section 5, conventional mediation estimands and separable effects are often identified by the same functionals of the observed data. The implication of this is that the existing computer software for conventional mediation estimands can often be used to calculate separable effects.

Moreover, the conditional separable effects are related to principal stratum effects.[15] Indeed, a conditional separable effect is restricted to subjects who have a certain value of the intercurrent event under the modified treatment a_S. Assuming that A_Y does not exert effects on S, (10) is equal to $\mathbb{E}(Y^{a_Y=1,a_S} - Y^{a_Y=0,a_S} \mid S^a = s)$ which targets the same subgroup considered in the PS estimand (6) rather than (7). However, it also explicitly quantifies a treatment effect that acts *outside* of the intercurrent event (a *direct* effect on the event of interest not mediated by the intercurrent event) in this subgroup. See Stensrud et al (15) for a discussion on the link between conditional separable effects and principal stratum estimands.

4 | TRANSLATING RESEARCH QUESTIONS TO ESTIMANDS

We now revisit the Examples 1 to 3 that were discussed in Section 2 (Examples 4 and 5 are discussed in Appendix B). We map research questions to their corresponding estimands. We find that all of these estimands are interventionist estimands that, at least in principle, could be studied in a (future) randomized trial.[13,44]

Example 1 (Multiple sclerosis (cont.))**.** In a secondary analysis of the EXPAND trial,[3] Magnusson et al[4] studied a principal stratum estimand like (6), but defined on the risk ratio scale,

$$\frac{\mathbb{E}(Y^{a=1} \mid S^{a=1} = S^{a=0} = 1)}{\mathbb{E}(Y^{a=0} \mid S^{a=1} = S^{a=0} = 1)}, \tag{11}$$

where Y indicates confirmed disability progression (at some time t) and the conditioning set $S^{a=1} = S^{a=0} = 1$ indicates having no relapses *regardless* of treatment assignment at baseline. Later this estimand was also advocated by the ASA/EFSPI oncology estimand working group.[1] Estimand (11) quantifies disability under siponimod vs no treatment in subjects who would not relapse under both siponoind and no treatment.[#]

One motivation for studying principal stratum effects, like (11), of siponimod was "*understanding the effect of siponimod on progression occurring independently of relapses*."[4] The relevance of effects in those who would not experience relapse regardless of treatment assignment is nevertheless unclear; as we cannot observe this subset of the population of unknown size, it cannot be a direct target population for a new drug in the future.

On the contrary, treatment effects *outside* of relapse are of interest if the investigators consider the opportunity of leveraging or avoiding these particular effects in future, refined drugs. For example, based on a biochemical evidence, the drug developers may have reasons to believe that the current drug exerts effects through different (biological) pathways, which have differential effects on relapse and disability progression. Questions about such mechanisms of action motivate the consideration of separable effects,[13,14,16,28,37,‖] which explicitly target the mechanism by which a treatment exerts effects on an outcome.[14-16,19,20,28,37] In particular, suppose that siponimod exerts effects on disability ($Y = 1$) through a pathway avoiding relapse ($S = 1$), and suppose that we could create a new drug ($A_Y = 1$) that exclusively targets this pathway, but does not act via the pathways by which siponimod reduces relapse in MS patients. Analogously, we could, in principle, imagine a different treatment ($A_S = 1$) that exclusively exerts effects on relapse, similar to siponimod, but does not exert effects on disability outside of the relapse pathway. This motivates separable effects estimands, defined with respect to a hypothetical trial where we assign the modified treatments A_Y and A_S,

$$\mathbb{E}(Y^{a_Y=1,a_S}) \text{ vs } \mathbb{E}(Y^{a_Y=0,a_S}), \tag{12}$$

which is defined (marginally) in the full study population. The contrast (12) quantifies the effect of siponimod vs the new drug that exclusively exerts effects on relapse. Thus, if the contrast (12) is equal to zero, then there is no effect of siponimod outside of its effect on relapse. If the contrast is different from zero, then siponimod also exerts effects outside of relapse on disability.

Magnusson et al[4] also state that there was particular interest in the treatment effect "*among the subgroup of patients for whom relapses would be absent during the study*." We may therefore consider a conditional separable effects estimand like (10),

$$\mathbb{E}(Y^{a_Y=1,a_S} \mid S^{a_S}=0) \text{ vs } \mathbb{E}(Y^{a_Y=0,a_S} \mid S^{a_S}=0).$$

This conditional separable effect quantifies the effect of siponimod vs the new drug that exclusively exerts effects on relapse, but, unlike (12), it is confined to those who would not relapse on siponimod.

So far we have discussed estimands that quantify effects through certain causal mechanisms, which, for example, can motivate the development of future drugs. However, if the aim is to support labeling decisions of siponimod itself, as is also suggested in Magnusson et al,[4] other estimands seem to be more relevant. In particular, doctors and regulators are often interested in whether the treatment effect varies across subgroups of patients. Such subgroup effects can only be useful for practical decision making if the subgroups are observed *before* the decision is made, which is not the case for principal stratum effects. On the other hand, simple average treatment effects conditional on a set of measured baseline covariates Z,

$$\mathbb{E}(Y^{a=1} - Y^{a=0} \mid Z=z), \tag{13}$$

is of immediate interest. Whereas the conditional average treatment effect is easy to define (and identify), summarizing subgroup effects across covariates $Z=z$ is not trivial. One way to summarize (coarsen) conditional effects is to define an auxiliary variable based on expected outcomes under treatment. In particular, instead of studying effects in principal strata, we could study effects for groups of patients who are *likely to be* in a principal stratum of interest.[45] Following Joffe et al,[45] define the principal score $\mu^a(Z)=P(S^a=1|Z)$,[46] which in our example denotes the probability of not having relapses under treatment $A=a$ given baseline covariates Z. The pair of principal scores $(\mu^{a=0}(Z), \mu^{a=1}(Z))$ is, unlike the principal stratum, a baseline covariate, because it is just a function of Z. We can now define causal effects among individuals with the same principal scores under treatment a, such as

$$\mathbb{E}(Y^{a=1} - Y^{a=0} \mid \mu^a(Z)=q), \text{ or} \tag{14}$$

$$\mathbb{E}(Y^{a=1} - Y^{a=0} \mid \mu^a(Z)>q), \tag{15}$$

which are the effect of siponimod on disability among patients with probability q or probability larger than q, respectively, of developing recurrences under treatment $A=a$. Similarly, we could study the joint principal scores, such as

$$\mathbb{E}(Y^{a=1} - Y^{a=0} \mid \mu^{a=1}(Z)=q_1, \mu^{a=0}(Z)=q_0), \tag{16}$$

which is the effect of siponimod on disability among patients with probability q_1 of developing recurrences when assigned to sifonimod and probability q_0 when assigned to placebo.

Unlike the principal stratum estimands, these principal score effects are identified at baseline. For example, a decision maker can be interested in giving a different treatment to patients with a high risk of relapses compared to patients with low risk of relapses. However, because the principal scores are just functions of Z, decision rules that use $\mu^{a=0}(Z)$ and/or $\mu^{a=1}(Z)$ as input will not be better than decision rules that only use Z as input.[**]

Example 2 (Treatment effects in early responders (cont.))**.** To motivate the study of treatment effects in early responders, Bornkamp et al[1] gave a subject-matter example on the effect of canakinumab (an immunotherapy) vs placebo on major adverse cardiovascular events (MACE), where the authors write that "*as the mechanism of action of canakinumab is lowering inflammation, one would suspect that patients who do not achieve the biomarker threshold also have a lower benefit in terms of the time-to-event outcome.*" The question concerns whether canakinumab *only* exerts effects on the outcome of interest through its effects on inflammation. Knowledge of this mechanism can motivate the development of new (refined) drugs, similarly to the motivating question in Example 1. To answer this question, a separable effect can be

explicitly formulated as follows: suppose that we could give a modified drug ($A_Y = 1$) in which the part of canakinumab that exerts effects on inflammation is blocked, but otherwise this drug is identical to canakinumab. Would this new drug have a beneficial effect on MACE compared to no treatment ($A_Y = 0$)?

If this new treatment has a beneficial effect, then there is a component of canakinumab that acts outside of inflammation. Specifically, this question corresponds to a separable effect (10), defined in the subset of the population who would not have a biomarker response *on* treatment (ie, those with $S^{a_S=1} = 0$). Unlike the joint principal stratum estimand

$$\mathbb{E}(Y^{a=1} \mid S^{a=1} = s) \text{ vs } \mathbb{E}(Y^{a=0} \mid S^{a=1} = s), \tag{17}$$

proposed in Bornkamp et al,[1] the separable effect explicitly quantifies a direct effect of treatment.

An alternative motivation for studying treatment effects in early responders is to "*support the decision on treatment modifications after treatment start.*"[1] This motivation does not concern drug development, but rather sequential decision making, and the substantive question motivates the study of a sequential trial estimand, as described in Section 3.1. To illustrate this point in the simplest possible setting, consider a trial in which treatment decisions can be made at two time points, $k = 0$ and $k = 1$, and suppose that the biomarker response is known at time $k = 1$ but not at time $k = 0$. Suppose further that patients can be assigned canakinumab ($A_t = 1$) or no treatment ($A_t = 0$) at times $k \in \{0, 1\}$. The authors' plain English motivation suggests the counterfactual estimand

$$\mathbb{E}(Y^{a_0=1,a_1=1} - Y^{a_0=1,a_1=0} \mid S^{a_0=1} = s). \tag{18}$$

This estimand quantifies the effect of a strategy that assigns canakinumab sequentially at baseline (time 0) and time 1 vs a strategy that assigns no treatment at baseline and time 1.

Example 3 (Impact of exposure on overall survival (cont.)). There is interest in "whether the lower OS is due to low drug concentration or to disease burden."[1,6,7] This causal question alludes to an effect of *low serum concentration of the active drug* vs no treatment on overall survival; for example, a practitioner may ask whether it is sufficient to have a low serum concentration to experience a treatment effect. Alternatively, the question may allude to the effect of *low* vs *high* serum concentration of the active drug on overall survival; for example, a practitioner may ask whether there is a dose-response effect. Both of these effects are specifically defined with respect to an intervention on the drug concentration itself: suppose, for example, that the variable $S = 1$ indicates that an individual has a drug concentration in the lowest quartile on treatment, and let $S = 0$ indicate that the drug concentration is equal to 0 (the concentration under no treatment). Then, the question posed by the investigators suggests the simple conditional effect

$$\mathbb{E}(Y^{s=1} - Y^{s=0} \mid X = x), \tag{19}$$

where X is a set of covariates that are available before the (hypothetical) decisions is made: here, X could be a trivial random variable if we are interested in the marginal effect of S, or X could be other variables temporary ordered before S, in particular it is possible that $A \subset X$. However, we emphasize that the exposure of interest in this setting is the drug concentration (S) in the blood, and not the drug administered (A). Indeed, a study conducted by researchers at the FDA[7] aimed to evaluate whether individuals with low serum concentration of the active drug would *show survival benefit* if they had a higher serum concentration.[7(p166)]

One reason for studying (19), which concerns drug concentration, is to assess whether a higher *dose* of the drug should preferably be administered, in particular to certain subgroups. In practice, however, these subgroups must be known before the treatment is given, which means that these subgroups *cannot* be principal strata (which are defined by intercurrent events that are unobserved at the time of treatment initiation). On the other hand, conditional treatment effects, possibly coarsened as principal scores like (14) and (15), could be used to make decisions; for example, it is possible to give a higher dose to those individuals who are likely to have a low serum concentration of the drug under a standard dose.

Indeed, after the FDA study,[7] a randomized trial has been conducted to evaluate the effect of different drug doses in certain groups of patients,[47] where these groups were solely defined by pretreatment variables, like X in (19).

5 | IDENTIFICATION ASSUMPTIONS

Our arguments in Sections 3 and 4 concerned the translation of a substantial research question, articulated in plain English, to a causal estimand, articulated in counterfactual notation. The arguments for choosing an estimand did not rely on the assumptions that are required to identify this estimand from the data at hand (eg, from an RCT where only A is randomly assigned). However, when an estimand of interest is chosen, it is crucial to understand, and critically assess, the assumptions necessary to learn about the estimand from data (ie, to evaluate whether a sufficient set of identifiability conditions hold). Furthermore, when setting up a randomized trial, these assumptions can be helpful in understanding how the study needs to be designed, or which data need to be collected.

Identifiability conditions for all of the estimands in Sections 3 and 4 are thoroughly described in several previous works.[15,21,28,44,48,49] Here we informally review some important features of these conditions, which should be justified on scientific grounds. We focus on nonparametric identifiability conditions. Sometimes (strong) parametric assumptions can be invoked to obtain alternative identification conditions; that is, identification conditions that only hold under a particular (parametric) model. For example, mixture models and Bayesian methods have often been used to identify principal stratum effects, but as noted by Bornkamp et al,[1] inference is then highly sensitive to the correctness of the parametric assumptions and likelihood estimators can exhibit pathological behavior. Indeed, the practical value of these alternative parametric conditions is limited unless the investigators have convincing arguments why we should believe in the parametric assumptions (which, to our knowledge, is rarely the case in medicine).

5.1 | The average total effect requires the weakest assumptions

Only the conventional average effect estimand (1) can be identified without additional assumptions from a (perfectly executed) RCT where the baseline treatment A is randomly assigned, as illustrated in the simple causal directed acyclic graph (DAG) in Figure 1A. This result is also valid for average effect that are defined conditional on observed baseline covariates, like the effect (19) and principal score effects like (14) and (15). All the other estimands we consider—including principal stratum effects, separable effects and sequential treatment effects—require additional assumptions in this setting, which we discuss in the following subsections.

5.2 | Sequential trial estimands

Sequential trial estimands can be identified without additional assumptions from a (perfectly executed) trial where the treatment is randomly assigned sequentially, that is, at multiple points in time. However, randomly assigning treatment at baseline (eg, in a randomized trial with treatment switching/discontinuation) is not sufficient to identify sequential trial estimands.[21,44,49] In particular, it is necessary to adjust for common causes of the outcome Y and the sequential treatment A_k at time k. Causal graphs allow us to visualize and evaluate the conditional independencies that ensure identification of sequential effects.[44,48] They are useful for translating and discussing the key assumptions with scientific collaborators, in order to assess their validity. For example, the causal DAG in Figure 1B describes a setting where the sequential trial estimands (5) and (18) are identified provided that S and Z_0 are measured.

5.3 | Separable effects require dismissible component conditions

The separable effects are defined with respect to modified treatments that, when combined, operate like the study treatments (see Section 3.3). These modified treatments may correspond to decompositions of the study treatments[13,28,37] or to treatments that are distinct from the study treatments but share the same mechanistic actions.[15,16,20] Like the sequential trial estimands, identification of separable effects requires conditional independencies (called dismissible component conditions or conditional exchangeabilities) to be satisfied.[13,14,16,28] Like the sequential trial estimands, these conditions require adjustment for common causes of the outcome Y and the intercurrent event S, which may be time-varying.[13,15,16,28] The dismissible component conditions would also be violated if certain variables are directly affected by both components A_Y and A_S.[††]

For example, if Z_0 is unmeasured in the DAG in Figure 1A, then the identification conditions for these estimands are violated. Variables like Z_0 are likely to be present in a range of practical settings. In Example 1, any factor (eg, related to genetic or lifestyle) that affects both relapse and disability would be classified as a Z_0. Unlike conventional mediation estimands, the separable effects are *single world* estimands that can be identified in a future randomized trial.[13,15,37] ‡‡

5.4 | Principal stratum estimands require alternative assumptions

Unlike the other estimands, identification of principal stratum effects require assumptions are empirically unverifiable, that is, untestable, even in principle.[15,28,30,44,51] In particular, identifying the subgroup of individuals that constitutes the principal stratum in (6) is impossible in most practical settings without relying on such unverifiable assumptions.[28,44] This limits the practical relevance of these estimands.[15,28-30,51] For example, a principal ignorability assumption is often invoked to identify principal stratum estimands such as (6) and (7).[52,53] Interpreting the principal ignorability assumption is not straight-forward: it is a cross-world independence assumption, which requires the investigator to reason about (untestable) independencies between counterfactuals under different treatment assignments. It is necessary (but not sufficient) that the investigator adjusts for common causes of the event of interest Y and the intercurrent event S, even in a study where A is randomly assigned, for principal ignorability to hold. These common causes may be time-varying, and, unlike the other estimands, the literature on principal stratum effects in the presence of time-varying confounding is scant and relies on assumptions that may be hard to justify.[15,26] On the other hand, principal score estimands, which are defined with respect to predicted probabilities of belonging to a stratum, can be identified under weaker conditions (see Section 5.1).

6 | A NOTE ON ESTIMATION AND SOFTWARE IMPLEMENTATION

This article focuses on interpretation and identification, not estimation. However, we would like to emphasize that our consideration of interpretation and identification has consequences for estimation. In particular, we can exploit the relations between the different estimands to clarify when the same estimation algorithms (and computer software) can be used to estimate different types of effects.

For example, computer software for natural direct and indirect effects can be used to estimate separable effects in any setting where they are identified by the same functionals of observed data distributions.[13,28,37] More generally, each separable effects can be estimated using software for path-specific effects; there is always a path-specific effect that is identified by the same functional as a separable effect.[13] See Valente et al[54] for a review of available software for causal mediation analysis.

Finally, many trials will be powered for the ITT analysis, and as such will not be guaranteed to be powered for the alternative estimands discussed in this article. Fang and Jin[55] proposed sample size calculation strategies for the estimands described in the ICH E9 addendum, and illustrated them using simulated data examples. Larger sample sizes are needed for all estimands relative to the ITT analysis; for example, a quarter increase is required for a principal stratification analysis. A major limitation of Fang and Jin's work is that post-treatment confounding is not considered. Confounding complicates power calculations, and may result in much larger sample sizes being needed to detect effects. In practice, some simulation studies based on plausible data generating mechanisms, possibly inspired by previous trial data, may help to give an impression of whether a trial is well-powered for the question of interest.

7 | DISCUSSION

We have clarified that certain causal estimands—including sequential treatment effects, separable effects and conditional causal effects—are useful in settings where principal stratum estimands are often recommended.[1,2] These estimands correspond to different causal questions. If the investigator is ultimately interested in finding subgroups where treatment works better, conditional causal effects are of interest, but principal stratum effects are not relevant because the principal strata are not observable when the decision is made. We illustrate this in Examples 1 to 3. Furthermore, separable effects

can disentangle how the treatment affects the outcome outside or through the intercurrent event, as we illustrate in Example 1 (and in Examples 4 and 5, which are discussed in Appendix B).

All of our suggested estimands share key features: they are defined in observable subsets of the population, they correspond to explicit causal inquiries, and they can at least in principle be empirically falsified in a future experiment.[13,15,37,44] These common features do not arise by coincidence. We believe that any estimand that guides practical decision making is in principle verifiable in a well-characterized population, because any practical decision can be instantiated in this population.

Causal inference is a subtle exercise, and being precise about the interpretation of estimands—and the practical questions that motivate them—is crucial to avoid logical flaws and erroneous decisions. Thus, we encourage investigators to be explicit about why they target a particular estimand, which itself is an exercise that can sharpen arguments and thought processes.[28,37]

When we have data from a trial where only a baseline treatment is randomly assigned, there are limitations as to how statistics can help us to discover causal mechanisms. Causal (structural) assumptions are required for this endeavor. These assumptions are not innocuous and are not guaranteed to hold. Estimates of separable effects can motivate the development of new treatment and generate new hypotheses of causal mechanisms, but the study of separable effects do not give guarantees that these components exist and can be used in future treatment. On the other hand, principal stratum effects do not guarantee that there exist (a substantial amount) of individuals that actually belongs to the principal stratum, do not allow us to characterize these individuals, and do not quantify the effect of future therapies.

If estimands beyond average total effects are of interest in a randomized clinical trial, which often seems to be the case, our work illustrates an important point: the investigator should strive to include a rich set of pre- and post-treatment variables. This is crucial, because identification of mechanistic estimands and sequential trial estimands would require adjustment of (possibly time-varying) common causes of the intercurrent event and the event of interest.

Finally, different estimands can, under certain assumptions, be identified by identical functionals of observed data distributions,[15] and therefore they can be estimated in the same way. Even if these estimands are identified by the same function of data, it is important that the investigators are explicit about their causal question and corresponding estimand: the interpretation of the estimand, and identification assumptions, can still differ,[14,15,28] which matters when these estimands are going to be used for practical purposes.

ACKNOWLEDGEMENTS

We thank Vanessa Didelez, Bjorn Bornkamp, Kaspar Rufibach, and Baldur Magnusson for generous comments and discussions that have helped us improve the article.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ENDNOTES

*Section 3 is a stand-alone section that can be passed over by readers who are already familiar with these estimands.

†Despite the fact that these interventions are easy to conceptualize, they may be impossible to implement in practice, for example, due to ethical reasons.

‡Unlike Frangakis and Rubin,[25] Robins[21] expressed a skeptical view of the practical importance of these estimands. The principal stratum effects in this survival setting are often denoted survivor average causal effects.[26,27]

§Unless additional assumptions are imposed.

¶For example, Robins and Richardson[28] discussed how Pearl made an argument based on modified treatments to motivate natural effects of cigarette smoking on cardiovascular disease. Stories about separable effects have also been (implicitly) been used to motivate mediation estimands in other settings, see, for example, References 38-40.

#We consider relapse at a time $s < t$ before assessment of the disability outcome $Y \equiv Y(t)$.

‖We use the term separable effect broadly to denote interventionists estimands for causal mechanisms,[13-16,28,37] which cover classical mediation settings, competing events and truncation by death.

**Here, "better" alludes to optimal expected outcomes $g^* \equiv \arg\max_{g \in \mathcal{G}} \mathbb{E}(Y^g)$, where g is a decision rule (regime).

††In this setting, a recanting witness makes it impossible to identify the separable effects.[13,15,28,37,50] However, note that recanting witnesses that preclude natural effects from being identified do not necessarily preclude a separable effect from being identified.[13,16,28]

‡‡Thus, identifiability conditions for separable effects can be evaluated in single world intervention graphs (SWIGs).[13,37,44]

§§As identification of principal stratum estimands require additional assumptions that are empirically untestable, even in principle, we cannot use SWIGs to study exchangeability conditions for principal stratum estimands.

¶¶Strictly speaking, this is a single world intervention template.[44]

ORCID

Mats J. Stensrud https://orcid.org/0000-0001-9641-1936

REFERENCES

1. Bornkamp B, Rufibach K, Lin J, et al. Principal stratum strategy: potential role in drug development. *Pharm Stat*. 2021.
2. ICH. Addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials E9(R1); 2019. https://databaseichorg.
3. Kappos L, Bar-Or A, Cree BA, et al. Siponimod versus placebo in secondary progressive multiple sclerosis (EXPAND): a double-blind, randomised, phase 3 study. *Lancet*. 2018;391(10127):1263-1273.
4. Magnusson BP, Schmidli H, Rouyrre N, Scharfstein DO. Bayesian inference for a principal stratum estimand to assess the treatment effect in a subgroup characterized by postrandomization event occurrence. *Stat Med*. 2019;38(23):4761-4771.
5. Ridker PM, Everett BM, Thuren T, et al. Antiinflammatory therapy with canakinumab for atherosclerotic disease. *N Engl J Med*. 2017;377(12):1119-1131.
6. Bang YJ, Van Cutsem E, Feyereislova A, et al. Trastuzumab in combination with chemotherapy versus chemotherapy alone for treatment of HER2-positive advanced gastric or gastro-oesophageal junction cancer (ToGA): a phase 3, open-label, randomised controlled trial. *Lancet*. 2010;376(9742):687-697.
7. Yang J, Zhao H, Garnett C, et al. The combination of exposure-response and case-control analyses in regulatory decision making. *J Clin Pharmacol*. 2013;53(2):160-166.
8. Cosson VF, Ng VW, Lehle M, Lum BL. Population pharmacokinetics and exposure–response analyses of trastuzumab in patients with advanced gastric or gastroesophageal junction cancer. *Cancer Chemother Pharmacol*. 2014;73(4):737-747.
9. Enrico D, Paci A, Chaput N, Karamouza E, Besse B. Antidrug antibodies against immune checkpoint blockers: impairment of drug efficacy or indication of immune activation? *Clin Cancer Res*. 2020;26(4):787-792.
10. Thompson IM, Goodman PJ, Tangen CM, et al. The influence of finasteride on the development of prostate cancer. *N Engl J Med*. 2003;349(3):215-224.
11. Michiels H, Sotto C, Vandebosch A, Vansteelandt S. A novel estimand to adjust for rescue treatment in clinical trials; 2020. arXiv preprint arXiv:200912052.
12. Follmann D. Augmented designs to assess immune response in vaccine trials. *Biometrics*. 2006;62(4):1161-1169.
13. Robins JM, Richardson TS, Shpitser I. An interventionist approach to mediation analysis; 2020. arXiv preprint arXiv:200806019.
14. Didelez V. Defining causal meditation with a longitudinal mediator and a survival outcome. *Lifetime Data Anal*. 2019;25:593-610.
15. Stensrud MJ, Robins JM, Sarvet A, Tchetgen EJT, Young JG. Conditional separable effects; 2020. *J Am Stat Assoc*. 2022 (accepted).
16. Stensrud MJ, Hernán MA, Tchetgen Tchetgen EJ, Robins JM, Didelez V, Young JG. A generalized theory of separable effects in competing event settings. *Lifetime Data Anal*. 2021;27(4):588-631.
17. Huang YT. Causal mediation of semicompeting risks. *Biometrics*. 2021. doi:10.1111/biom.13525
18. Fulcher IR, Shpitser I, Didelez V, Zhou K, Scharfstein DO. Discussion on ``Causal mediation of semicompeting risks'' by Yen-Tsung Huang. *Biometrics*. 2021.
19. Stensrud MJ, Young JG, Martinussen T. Discussion on ``Causal mediation of semicompeting risks'' by Yen-Tsung Huang. *Biometrics*. 2021.
20. Young JG, Stensrud MJ. Identified versus interesting causal effects in fertility trials and other settings with competing or truncation events. *Epidemiology*. 2021.
21. Robins JM. A new approach to causal inference in mortality studies with a sustained exposure period—Application to control of the healthy worker survivor effect. *Math Model*. 1986;7(9-12):1393-1512.
22. Hernan MA, Robins JM. *Causal Inference*. Boca Raton, FL: CRC Press; 2018.
23. Hernán MA, Scharfstein D. Cautions as regulators move to end exclusive reliance on intention to treat. *Am Coll Physicians*. 2018;168:515-516.
24. Hernán MA, Robins JM. Per-protocol analyses of pragmatic trials. *N Engl J Med*. 2017;377(14):1391-1398.
25. Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics*. 2002;58(1):21-29.
26. Tchetgen Tchetgen EJ. Identification and estimation of survivor average causal effects. *Stat Med*. 2014;33(21):3601-3628.
27. Egleston BL, Scharfstein DO, MacKenzie E. On estimation of the survivor average causal effect in observational studies when important confounders are missing due to death. *Biometrics*. 2009;65(2):497-504.
28. Robins JM, Richardson TS. Alternative graphical causal models and the identification of direct effects. *Causality and Psychopathology: Finding the Determinants of Disorders and Their Cures*; 2010:84:103-158.
29. Robins J, Rotnitzky A, Vansteelandt S, Have TT, Xie Y, Murphy S. Discussions on principal stratification designs to estimate input data missing due to death. *Biometrics*. 2007;63(3):650-658.
30. Joffe M. Principal stratification and attribution prohibition: good ideas taken too far. *Int J Biostat*. 2011;7(1):35.
31. Dawid P, Didelez V. Imagine a can opener–The magic of principal stratum analysis. *Int J Biostat*. 2012;8(1):19.
32. VanderWeele TJ. Principal stratification—Uses and limitations. *Int J Biostat*. 2011;7(1):28.
33. Pearl J. Principal stratification—A goal or a tool? *Int J Biostat*. 2011;7(1):1-13.
34. Luedtke A, Wu J. Efficient principally stratified treatment effect estimation in crossover studies with absorbent binary endpoints; 2017. arXiv preprint arXiv:171205835.

35. Wolfson J, Gilbert P. Statistical identifiability and the surrogate endpoint problem, with application to vaccine trials. *Biometrics*. 2010;66(4):1153-1161.
36. Stensrud MJ, Smith LH. Identification of vaccine effects when exposure status is unknown; 2021. arXiv preprint arXiv:211111548.
37. Stensrud MJ, Young JG, Didelez V, Robins JM, Hernán MA. Separable effects for causal inference in the presence of competing events. *J Am Stat Assoc*. 2020;1-23.
38. VanderWeele TJ. Explanation in causal inference: developments in mediation and interaction. *Int J Epidemiol*. 2016;45(6):1904-1908.
39. Lange T, Hansen KW, Sørensen R, Galatius S. Applied mediation analyses: a review and tutorial. *Epidemiol Health*. 2017;39:e2017035.
40. Wilkinson J, Huang JY, Marsden A, Harhay MO, Vail A, Roberts SA. The implications of outcome truncation in reproductive medicine RCTs: a simulation platform for trialists and simulation study. *Trials*. 2021;22(1):1-15.
41. Buettner C, Davis RB, Leveille SG, Mittleman MA, Mukamal KJ. Prevalence of musculoskeletal pain and statin use. *J Gen Intern Med*. 2008;23(8):1182-1186.
42. Nissen SE, Stroes E, Dent-Acosta RE, et al. Efficacy and tolerability of evolocumab vs ezetimibe in patients with muscle-related statin intolerance: the GAUSS-3 randomized clinical trial. *JAMA*. 2016;315(15):1580-1590.
43. Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*. 1992;3(2):143-155.
44. Richardson TS, Robins JM. Single world intervention graphs (SWIGs): a unification of the counterfactual and graphical approaches to causality. *Center Stat Soc Sci Univ Washington Ser Working Pap*. 2013;128(30):2013.
45. Joffe MM, Small D, Hsu CY, et al. Defining and estimating intervention effects for groups that will develop an auxiliary outcome. *Stat Sci*. 2007;22(1):74-97.
46. Hill J, Waldfogel J, Brooks-Gunn J. Differential effects of high-quality child care. *J Policy Anal Manag J Assoc Public Policy Anal Manag*. 2002;21(4):601-627.
47. Shah MA, Xu R, Bang YJ, et al. HELOISE: Phase IIIb randomized multicenter study comparing standard-of-care and higher-dose trastuzumab regimens combined with chemotherapy as first-line therapy in patients with human epidermal growth factor receptor 2–positive metastatic gastric or gastroesophageal junction adenocarcinoma. *J Clin Oncol*. 2017;35(22):2558-2567.
48. Pearl J. *Causality: Models, Reasoning and Inference*. 2nd ed. Cambridge, UK: Cambridge University Press; 2000.
49. Hernán MA, Robins JM. *Causal Inference: What If*. Boca Raton, FL: Chapman & Hill/CRC Press; 2020:2020.
50. Shpitser I. Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cogn Sci*. 2013;37(6):1011-1035.
51. Dawid P, Didelez V. Identifying the consequences of dynamic treatment strategies: a decision-theoretic overview. *Stat Surv*. 2010;4:184-231.
52. Jo B, Stuart EA. On the use of propensity scores in principal causal effect estimation. *Stat Med*. 2009;28(23):2857-2875.
53. Ding P, Lu J. Principal stratification analysis using principal scores. *J Royal Stat Soc Ser B (Stat Methodol)*. 2017;79(3):757-777.
54. Valente MJ, Rijnhart JJ, Smyth HL, Muniz FB, MacKinnon DP. Causal mediation programs in R, M plus, SAS, SPSS, and stata. *Struct Equ Model Multidiscip J*. 2020;27(6):975-984.
55. Fang Y, Jin M. Sample size calculation when planning clinical trials with intercurrent events. *Ther Innov Regul Sci*. 2021;55(4):779-785.

How to cite this article: Stensrud MJ, Dukes O. Translating questions to estimands in randomized clinical trials with intercurrent events. *Statistics in Medicine*. 2022;41(16):3211-3228. doi: 10.1002/sim.9398

APPENDIX A. MORE DETAILS ON SEPARABLE EFFECTS

The separable effects are defined with respect to modified treatment components, A_Y and A_S, which are linked to the original treatment A in the following way: when A_Y and A_S are set to the same value a, that is $A_Y = A_S = a$, then the counterfactual values of the event of interest $Y^{a_Y=a,a_S=a}$ and the intercurrent event $S^{a_Y=a,a_S=a}$ are equal to the counterfactual outcomes when the original treatment $A = a$, that is, $Y^{a_Y=a,a_S=a} = Y^a$ and $S^{a_Y=a,a_S=a} = S^a$. This modified treatment assumption holds when the original treatment A can be decomposed into two components A_Y and A_S. However, this assumption can also hold even if A cannot be physically decomposed.[15] When the components A_Y and A_S are given individually, they exclusively target certain causal pathways. For example, let A be statin therapy and Y be cardiovascular risk. The combination $A_Y = 1$ and $A_S = 0$ can be conceived as a modified therapy (such as PCSK9 inhibitors) that, like statins, have a cholesterol lowering component $A_Y = 1$, but lacks effects on the intercurrent event such as muscle symptoms, $A_S = 0$. To study separable effects, the pathways by which A_Y and A_S exert effects must be isolated from each other; in particular, we require that either A_Y component exerts its effects in Y or the A_S component exerts effects its on S.

The pathways by which A_Y and A_S exert effects can be described by isolation conditions,[15,16,37] which can be expressed as follows:

$$\text{There are no causal paths from } A_Y \text{ to } S, \text{and} \tag{A1}$$

$$\text{there are no causal paths from } A_S \text{ to } Y. \tag{A2}$$

Conditions (A1) and (A2) are called A_Y partial isolation and A_S partial isolation, respectively. These assumptions can be generalized to time-varying settings;[13,15,16] like Bornkamp et al,[1] here we consider (simplified) intercurrent event setting, where there is a single outcome Y (possibly a survival outcome) and a single intercurrent event S.

To define separable effects, we consider (hypothetical) settings where the components A_Y and A_S are assigned separately, and thus can be given different values. More precisely, we define the separable effect of the A_Y component as

$$\Pr(Y^{a_Y=1,a_S}=1) \text{ vs } \Pr(Y^{a_Y=0,a_S}=1), \quad a_S \in \{0,1\}, \tag{A3}$$

which quantifies the causal effect of the A_Y component on the risk of the outcome of under an intervention that assigns $A_S = a_S$. Similarly,

$$\Pr(Y^{a_Y,a_S=1}=1) \text{ vs } \Pr(Y^{a_Y,a_S=0}=1), \quad a_Y \in \{0,1\}, \tag{A4}$$

quantifies the causal effect of the A_S component on the event of interest under an intervention that assigns $A_Y = a_Y$.

Under (A1) and (A2) the separable effect (A3) quantifies the direct effect of the treatment on the outcome of interest, but this direct effect is different from the principal stratum estimand because it is defined in the entire study population. Similarly, under (A1) and (A2) the separable effect (A4) quantifies the indirect effect of the treatment on the outcome of interest through the intercurrent event.

Like conventional mediation estimands,[13,28] such as natural direct and indirect effects,[43] the separable effect quantify the mechanisms by which the treatment affects the outcome of interest.[13,14,16,28,37] Unlike the conventional mediation estimands, the separable effects do not require specification of any intervention on the intermediate event. It is not clear that such interventions on the intercurrent events exist in any of Examples 1 to 5. Furthermore, a feature of the separable effects is that they are defined with respect to modified treatments, and therefore they can be directly relevant in a drug development setting where the interest is tailoring new treatments to include/exclude certain causal pathways.

In settings with intercurrent events, researchers often express interest in estimands *conditional* on the intercurrent events. Following Stensrud et al,[15] under the assumption that A_Y partial isolation (A1) holds, we can also define a conditional separable effect as

$$\mathbb{E}(Y^{a_Y=1,a_S} - Y^{a_Y=0,a_S} \mid S^{a_S} = s).$$

The conditional separable effect is the average causal effect of the (modified) treatment A_Y on Y when all individuals are assigned the other (modified) treatment $A_S = a_S$ in those individuals who do not experience the intercurrent event under $A_S = a_S$ (regardless of their value of A_Y).

The conditional separable effect is restricted to subjects who have a certain value of the intercurrent event under modified treatment a_S. Under the partial isolation assumption, (10) is equal to $\mathbb{E}(Y^{a_Y=1,a_S} - Y^{a_Y=0,a_S} | S^a = s)$ which targets the same subgroup considered in the PS estimand (6) rather than (7). However, it also explicitly quantifies a treatment effect that acts *outside* of the intercurrent event (a *direct* effect on the event of interest not mediated by the intercurrent event) in this subgroup.

APPENDIX B. FURTHER ELABORATION ON EXAMPLES 4 AND 5

Example 4 (ADA for targeted oncology trials (cont.))**.** There is interest in whether ADAs bind to a *region* of the immunotherapeutic drug that affects its efficacy. In other words, there is interest in the mechanisms by which the immunotherapy exerts effects on the clinical outcome (say, all-cause survival at a given time t) in the presence of ADAs.[1,5]

Hence, despite the fact that survival is a well-defined outcome whether or not the intercurrent event (ADA production) occurs, a study of total treatment effects on survival will not provide information on the effect of treatment in the presence of ADAs.

Nevertheless, the story about mechanisms immediately motivates a study of a separable effect, similarly to Example 1 and the first question in Example 2. The authors also concretely suggest (physical) components of treatment[14-16,28,37] that exert different effects: Let A indicate immunotherapy ($A = 1$) or no treatment ($A = 0$). The immunotherapy ($A = 1$) can be decomposed into a component to which the antibody binds ($A_S = 0$, a particular part of the drug), and the remaining component to which it does not bind (say, $A_Y = 1$, the remaining part of the drug). Then, the estimand in (10), conditional on $S^{a_S=1} = S^{a=1} = s$ (those who would have ADA production when treated), would quantify the effect of the component to which the antibody binds (A_Y) on the clinical outcome (Y, say, survival after a time t) among those who would get ADA under treatment. Assuming that nobody can develop ADAs without receiving treatment (a monotonicity assumption) and full isolation,[15] the conditioning set defined by the separable effects will coincide with an estimand defined in the union of principal strata by Bornkamp et al.[1] Yet, the motivation for choosing the principal stratum estimand is not clear unless a causal question is explicitly stated and translated to a formal estimand: the theory of separable effects is precisely created to help with this task,[15,28,37] and the reasoning about separable effects is precisely what justifies why the estimand defined in a union of principal stratum is of substantial interest. In other words, the separable effects reasoning is required to justify the use of the principal stratum estimand.

Example 5. (Prostate cancer prevention (cont.))**.** The investigators expressed interest in "the effect of finasteride on the severity of prostate cancer among those men who would be diagnosed with prostate cancer regardless of their treatment assignment."[1] This sentence alludes to a principal stratum estimand, similar to the setting in Examples 1,2, and 4. However, if the research question is whether finasteride does not just prevent mild cancers but also leads to more aggressive cancer via a separate mechanism, this calls for a different estimand. If finasteride exerted such harmful effects, drug developers could aim to modify finasteride (or construct an alternative treatment), such that these harmful effects were avoided while preserving the preventive effect. Again, this motivates a separable effect, corresponding to estimand (8): Let $A_S = 1$ be a new drug that exerts the same effects as finasteride on cancer prevention, and let $A_Y = 1$ be another (potentially harmful) substance that has the same effect as finasteride on everything else except cancer progression. The question of the decision maker seems to be whether a new drug, which exerts the effect of the component $A_S = 1$, but does not exert any of the other effects of finasteride (that is, $A_Y = 0$), would be better than administering finasteride (now we can think of finasteride as equivalent to a drug containing the components $A_Y = 1, A_S = 1$).

APPENDIX C. SINGLE WORLD INTERVENTION GRAPHS

It is possible to draw single world intervention graphs (SWIGs)[44] to evaluate exchangeability conditions for all interventionist estimands,[13] such as sequential trial estimands and separable effects. SWIGs are causal graphs that are related to conventional causal DAGs.[48] Unlike a conventional DAG, the SWIG explicitly encodes a setting where the intervention of interest is instantiated (see Reference 44) and therefore allows us to directly evaluate counterfactual independencies for the estimand of interest. A feature of SWIGs is that these graphs only encode assumptions that are empirically falsifiable, that is, testable, in a (future) study.[§§]Splitting of nodes in a SWIG describes interventions on the nodes (ie, variables) of interest, and these interventions define the counterfactual estimand. For example, the simple SWIG in Figure 2A describes a setting where we consider counterfactual outcomes (Y^{a_Y,a_S}) under interventions on the treatment components A_Y and A_S. This graph can be used to read off whether *counterfactual* independencies between the treatment components A_Y and A_S and the counterfactual outcome Y^{a_Y,a_S} hold.[¶¶]The SWIGs can also be used to evaluate identification conditions for sequential trial estimands such as (5) in Example 4, as illustrated by the SWIG in Figure 2B. Similarly, SWIGs can be used to evaluate whether the effect of S in Example 3 can be identified, which require standard conditions for identification of (conditional) causal effects.[49] In particular, the SWIG in Figure 2C describes a setting where (19) is identified if we observe A and Z_0.

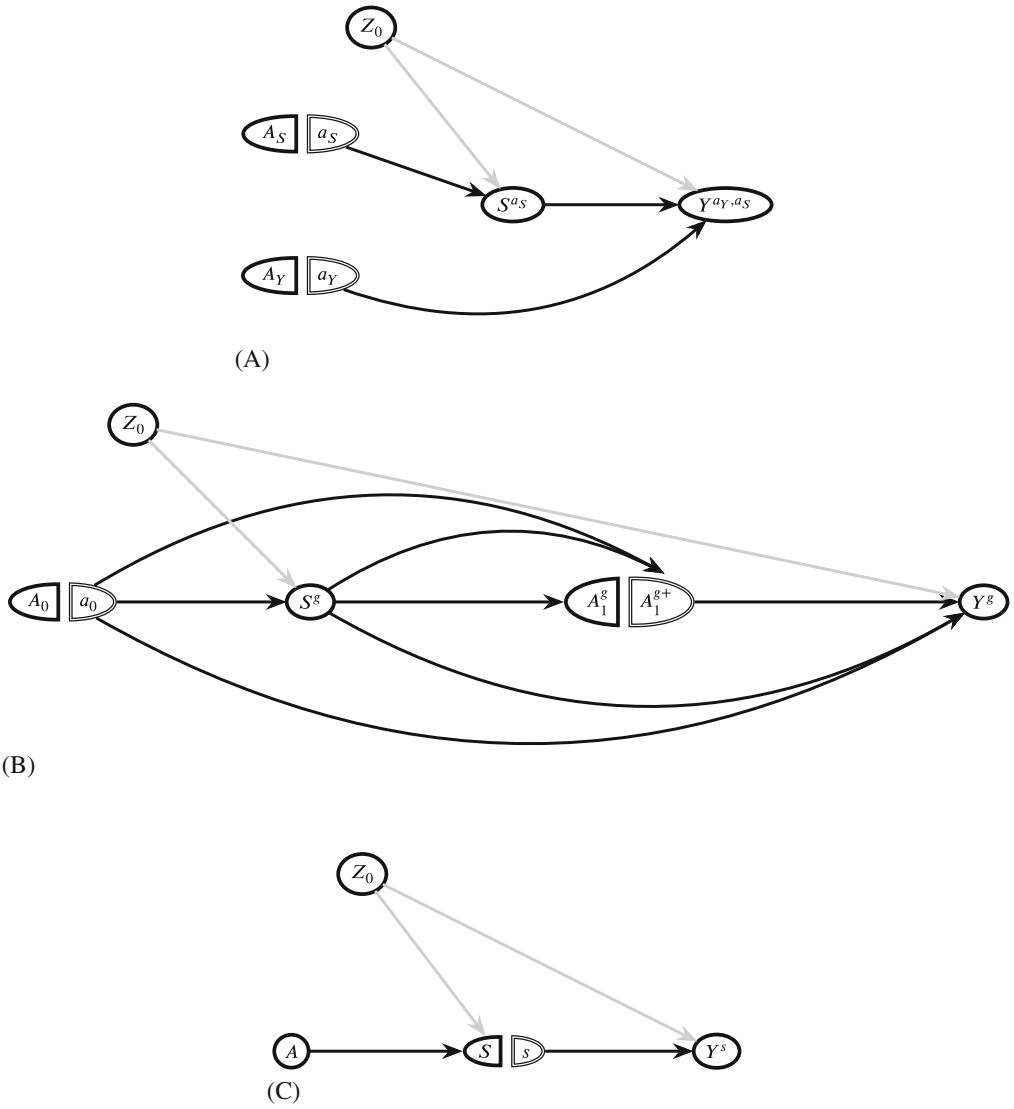

FIGURE C1 (A-C) Single world intervention graphs (SWIGs) with minimal labeling.[44] Superscripts denote counterfactuals. The SWIG in (A) describes a setting where conditional separable effects are identified, provided that Z_0 is measured. The SWIG in (B) describes a more involved setting The SWIGs are minimal labeled. The SWIG in (C) describes a setting where the dynamic treatment regime g is identified conditional on Z_0