Conditional Separable Effects

Mats J. Stensrud\textsuperscript{a}, James M. Robins\textsuperscript{b,c}, Aaron Sarvet\textsuperscript{a}, Eric J. Tchetgen Tchetgen\textsuperscript{d}, and Jessica G. Young\textsuperscript{b,c,e}

\textsuperscript{a}Department of Mathematics, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland; \textsuperscript{b}Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA; \textsuperscript{c}CAUSALab, Harvard T. H. Chan School of Public Health, Boston, MA; \textsuperscript{d}Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA; \textsuperscript{e}Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA

\begin{abstract}
Researchers are often interested in treatment effects on outcomes that are only defined conditional on posttreatment events. For example, in a study of the effect of different cancer treatments on quality of life at end of follow-up, the quality of life of individuals who die during the study is undefined. In these settings, naive contrasts of outcomes conditional on posttreatment events are not average causal effects, even in randomized experiments. Therefore, the effect in the principal stratum of those who would have the same value of the posttreatment variable regardless of treatment (such as the survivor average causal effect) is often advocated for causal inference. While principal stratum effects are average causal effects, they refer to a subset of the population that cannot be observed and may not exist. Therefore, it is not clear how these effects inform decisions or policies. Here we propose the conditional separable effects, quantifying causal effects of modified versions of the study treatment in an observable subset of the population. These effects, which may quantify direct effects of the study treatment, require transparent reasoning about candidate modified treatments and their mechanisms. We provide identifying conditions and various estimators of these effects along with an applied example. Supplementary materials for this article are available online.
\end{abstract}

1. Introduction

Many research questions involve treatment effects on outcomes that are only defined conditional on a posttreatment event status. For example, in a study of the effects of different cancer treatments on quality of life at end of follow-up, the quality of life of individuals who die during the study is undefined. Furthermore, some treatment effects are only of substantive interest conditional on a posttreatment event. In a study of the effect of a vaccine on viral load at end of follow-up, the viral load of individuals who never become infected is not of substantive interest (even though it is defined).

By design of a randomized trial, we can identify a counterfactual contrast of mean outcomes across treatment arms conditional on a posttreatment status (when there are no losses to follow-up). However, this contrast does not in general equal a causal effect when the treatment affects the posttreatment event. In this case, outcomes are being compared in different sets of individuals, and thus the comparison cannot be interpreted as a contrast of expected (counterfactual) outcomes in the same set of individuals under different treatment conditions. For example, when the cancer treatments affect survival, the subset of the population who would survive under one treatment will be different from the subset who would survive under the other treatment.

Selecting a meaningful definition of a causal effect in this setting is not straightforward. One option is to consider a so-called controlled direct effect (Robins and Greenland 1992), which quantifies the effect of the treatment on the outcome had we (somehow) eliminated posttreatment events that render the outcome undefined or not of substantive interest. However, such effects often do not quantify effects of interest to an investigator, policy maker, doctor or patient; for example, the utility of the effect of a cancer treatment on quality of life had we eliminated death, or the effect of a vaccine on viral load had we forced everyone to become infected, is not clear.

Given the limitations of the controlled direct effect, Robins (1986) introduced the principal stratum effects in settings where the outcome of interest is only defined, or of substantive interest, conditional on a posttreatment event status; in particular, the causal effect in the subset of individuals who would have this event status, regardless of the treatment they were given. The name “principal stratum effect” is due to Frangakis and Rubin (2002) who, in contrast to Robins’s (1986) more skeptical view, advocated strongly for the use of this estimand. Indeed, it has been argued that no other sensible causal estimand exists in this case (Rubin 2006; VanderWeele 2011; Ding and Lu 2017). However, principal stratum effects also have several serious limitations (Robins 1986; Robins, Rotnitzky, and Vansteelandt 2007; Joffe 2011; Robins and Richardson 2010; Dawid and Didelez 2012; Robins, Richardson, and Shpitser 2020). It is impossible to observe who comprises this subset of the population because it is defined by the event status in the same individual under different treatments. Further, this subset may not actually exist and, if it does exist, may constitute a highly unusual subset of
the original population. Therefore, it is unclear how a principal stratum effect informs treatment decisions or policies (i.e., how they inform actions). A consequence of this is that identification of a principal stratum effect generally relies on assumptions that cannot be falsified in any real-world experiment.

Here we provide definitions of new causal effects for settings where the outcome is only defined or of substantive interest conditional on a particular posttreatment event status: the conditional separable effects. We give general conditions under which these effects can be identified along with various estimators. These estimands are inspired by Robins and Richardson's treatment decomposition in the context of mediation (Robins and Richardson 2010; Robins, Richardson, and Shpitser 2020) and the (marginal) separable effects for competing events settings (Stensrud et al. 2021, 2022). Both marginal and conditional separable effects constitute effects of modified versions of the current study treatment with particular mechanisms removed, requiring transparent reasoning about candidates for modified treatments and their mechanisms. However, the problem of identifying the conditional separable effects is nontrivially distinct from the problem of identifying their marginal counterparts, leading to distinct interpretations and identifying functionals from those considered in this previous work. In turn, the estimation problem here is distinct and the estimators we develop do not overlap with those developed in this previous work.

The conditional separable effects rely on a key isolation condition which ensures that (i) these effects quantify direct effects of the treatment on the outcome in a particular subset of the population and (ii) the individuals comprising this subset can be identified. Regardless of whether the isolation condition holds, we argue that critical thinking about this condition is essential when the outcome is undefined or not of substantive interest on a posttreatment event status. In particular, we argue that when investigators cannot explicitly articulate modified treatments that meet this condition, it is not clear that there currently exists a notion of the treatment effect on this outcome that is useful to decision makers. This view is reinforced by numerous published examples of authors who explicitly construct analyses for principal stratum effects or controlled direct effects under inconceivable interventions on posttreatment events, yet ultimately justify the clinical or public health relevance of these analyses by discussing a different effect—an effect of a modified version of the study treatment with particular mechanisms removed (see, e.g., VanderWeele 2016, p. 1907; Lange et al. 2017, p. 2; Wilkinson et al. 2020, p. 25; Bornkamp et al. 2021). Finally, we show that, given isolation conditions, a particular conditional separable effect equals a principal stratum effect under an additional monotonicity condition.

The manuscript is organized as follows. In Section 2 we describe the observed data structure. In Section 3, we review the interpretation and identification of conventional principal stratum effects. In Section 4, we define the conditional separable effects and discuss their interpretation under the isolation condition. In Section 5, we define a modified treatment assumption (Stensrud et al. 2021), allowing the conditional separable effects to explain the mechanism of the original treatment under study. In Section 6 we discuss the need for the isolation condition and modified treatment assumption in order to define a meaningful effect of the original treatment on the outcome in this setting and give an example. In Section 7, we define conditions that are sufficient to identify the conditional separable effects with the observed data, provide an identification formula, and give new identification results for classical principal stratum estimands in the presence of time-varying common causes of the outcome of interest and the conditioning event. In Section 8, we derive and describe different estimators of conditional separable effects, including a doubly robust estimator based on the nonparametric influence function. In Section 9, we apply our results to understand the effect of different chemotherapies on quality of life in patients with prostate cancer. In Section 10, we end with a discussion.

2. Observed Data

Consider a randomized experiment with iid individuals who are assigned a binary treatment \( A \in \{0, 1\} \) at baseline. Let \( k \in \{0, \ldots, K + 1\} \) index equally spaced discrete time intervals, and \( Y \equiv Y_{k+1} \) be the outcome of interest measured at the final time \( K + 1 \). Let \( D_{k+1} \) be an indicator of posttreatment event status by \( k + 1 \), that is, \( D_{k+1} \) is nondecreasing in \( k \), such that \( Y \) is only defined, or is only of substantial interest, when \( D_{k+1} = 0, k \in \{0, \ldots, K\} \); otherwise, when \( D_{K+1} = 1, Y \) is undefined/not of substantive interest (\( Y \) can be understood to take a meaningless value in this case, e.g., \( Y = \ast \)). In our running example on cancer treatment and quality of life, \( D_{k+1} \) is an indicator of death by \( k + 1 \). In our other example on vaccination and viral load, \( D_{k+1} \) is an indicator of not having been infected by \( k + 1 \).

Let \( L_0 \) denote a vector of measured prerandomization (baseline) covariates, and \( L_k \) a vector of postrandomization (time-varying) covariates measured in interval \( k \).\(^1\) We adopt the temporal convention \((L_0, A, \ldots, D_k, L_K, \ldots, D_{K+1}, Y)\), which does not impose restrictions on the distribution of observed data when we let the interval length become infinitesimally small. We use overbars to denote the history of a random variable through \( k \), for example, \( \bar{D}_k \equiv (D_0, \ldots, D_k) \), and underbars to denote its future relative to \( k + 1 \), for example, \( \underline{D}_{k+1} \equiv (D_{k+1}, \ldots, D_{K+1}) \). We assume that no subject is lost to follow-up throughout the main text but extend all results to settings with loss to follow-up (censoring) in supplementary material A.

3. Challenges in Interpretation and Identification of Principal Stratum Effects

For any individual in the study population, let the counterfactual variables (alternatively referred to as potential variables) \( Y^a \) and \( D_{k+1}^a, k \in \{0, \ldots, K\} \), denote the outcome of interest and

\(^1\)In this presentation, we intentionally distinguish notation for time-varying covariates that are measured in our study, \( L_k \), from time-varying covariates that may or may not be measured but impact isolation conditions and interpretation of counterfactual estimands. We will need to make assumptions about the nature of \( L_k \) to reason about whether separable effects can be identified using only what was measured in our study. As in our previous work (Stensrud et al. 2021), we keep the tasks of interpretation and identification separate because explicit reasoning about interpretation of separable effects provides value for the design of future studies even if identification in the current study fails given limitations of measurement. This may be the case if causal reasoning about questions and assumptions occurs after the data collection is complete.
the post-treatment event indicator by \( k + 1 \), respectively, had, possibly contrary to fact, she been assigned to \( A = a \). The additive principal effect of the treatment \( A \) on \( Y \) in the subset of individuals who would never experience the posttreatment event rendering \( Y \) undefined/not of interest under any level of treatment (Robins 1986; Frangakis and Rubin 2002). Because the conditioning set in (1) is defined by outcomes in the same individual but under different treatments, it is impossible to observe the individuals in this subset of the population (Robins 1986), which clearly limits its practical relevance (Robins, Rotnitzky, and Vansteelandt 2007; Joffe 2011; Dawid and Didelez 2012). Further, we are not guaranteed that this unknown subset of the population exists and, if it does exist, it may constitute a highly unusual subset of the original study population.

Returning to our cancer treatment example, (1) is the effect of cancer treatment on quality of life at end of follow-up among those who would survive throughout the study regardless of what cancer treatment they received. This may be a highly unusual subgroup, particularly if \( a = 0 \) refers to no treatment. Similarly, in our vaccine example, (1) is the effect of receiving the vaccine on viral load at end of follow-up among those who would become infected regardless of whether they received the vaccine. For an effective vaccine, this subgroup may constitute a small and unusual segment of the original population (which, again, cannot be observed).

A consequence of these interpretational challenges is that strong assumptions are required for identification of principal stratum effects like (1) (Robins 1986; Hayden, Pauler, and Schoenfeld 2005; Robins, Rotnitzky, and Vansteelandt 2007; Tchetgen Tchetgen 2014; Ding and Lu 2017), even in idealized settings with a randomly assigned point treatment, no loss to follow-up and no common causes of \( Y \) and \( D_t \) as represented in the causal directed acyclic graph (DAG) (Pearl 2000) in Figure 1(a). Throughout, we will use causal DAGs to represent underlying assumptions on how random variables in a particular study are generated. Specifically, we use causal DAGs to represent an underlying Finest Fully Randomized Causally Interpreted Structural Tree Graph (FFRCISTG) model (as fine as the data) (Robins 1986; Richardson and Robins 2013), which is a counterfactual causal model that predates and makes fewer assumptions than the perhaps more familiar nonparametric structural equation model with independent errors (NPSEM-IE) (Pearl 2000; Robins and Richardson 2010; Richardson and Robins 2013). The absence of an arrow on a causal DAG representing a FFRCISTG model can either encode (i) the assumption that a population level causal effect is absent (Dawid 2000; Richardson and Robins 2013; Dawid 2015; Robins, Richardson, and Shpitser 2020). All of our results with respect to identification and estimation remain valid under the weaker population level assumption, also under an agnostic causal model that does not postulate the existence of counterfactual outcomes (Dawid 2000, 2015; Richardson and Robins 2013). However, the distinction between the individual level versus population level interpretation of an FFRCISTG will impact the interpretation of counterfactual contrasts introduced below, as we will note where relevant. Furthermore, in contrast to the NPSEM-IE model, the FFRCISTG model only requires counterfactuals to be defined with respect to interventions on the treatment variable(s) under consideration.\(^2\)

A causal DAG must minimally represent all common causes of any variable represented on the DAG. Therefore, the DAG in Figure 1(a) represents a restrictive assumption on the study data generating process because it depicts no common causes (measured or unmeasured) of \( D_t \) and \( Y \), which cannot be guaranteed even in a perfectly executed trial. Only a handful of authors have considered identification of principal stratum effects when common causes of the conditioning event and the outcome of interest exist and may be affected by treatment (Tchetgen Tchetgen 2014); this more realistic data generating assumption is represented in Figure 2(a), which, unlike the graph of Figure 1(a), includes the event \( D_t \) at two times \( k \in \{1, 2\} \) and a time-varying common cause \( Z_{k-1} \) of \( D_k \) and \( Y \). Further, previously posed identification strategies for (1) have relied on un falsifiable assumptions; that is, assumptions that can never be challenged in any possible experiment.

\(^2\)In this article we only consider counterfactuals associated with interventions on the treatment, the modified treatments introduced in Section 4, and loss to follow-up (see supplementary material A).
In the sections that follow, we will introduce new counterfactual estimands that overcome these limitations of principal stratum effects. We will formalize assumptions under which we can identify these new estimands, allowing common causes of the conditioning event and the outcome of interest affected by treatment. Finally, we show that, under an additional monotonicity assumption, a case of these estimands coincides with a principal stratum effect such as \((1)\), thus, providing new identifiability assumptions for modified treatments \(A_Y\) and \(A_D\), where we have omitted \((G)\) in each node to avoid clutter. Full isolation holds in (b), only \(A_Y\) partial isolation holds in (c) and only \(A_D\) partial isolation holds in (d).

4. The Conditional Separable Effects

Following Robins and Richardson (2010) and Robins, Richardson, and Shpitser (2020) in a mediation context and Stensrud et al. (2021, 2022) in a competing event context, suppose a four-arm trial could be plausibly conducted such that, in place of assignment to one of the two values of \(A\) as in Section 2, individuals are jointly assigned values of two new treatments \(A_Y \in \{0, 1\}\) and \(A_D \in \{0, 1\}\). In this section, we will study features of these new treatments \(A_Y\) and \(A_D\), and we will return to the relation between \(A_Y\), \(A_D\) and \(A\) in Section 5. Let \(Y^{a_Y, a_D}\) and \(D_{k+1}^{a_Y, a_D}\) denote the counterfactual outcome of interest and the posttreatment event indicator by \(k+1\), respectively, had an individual possibly contrary to fact been assigned \(A_Y = a_Y\) and \(A_D = a_D\) for \(a_Y\) and \(a_D\) possible realizations of \(A_Y\) and \(A_D\), respectively. We denote this four-arm trial by \(G\). Consider the following condition relative to a causal DAG representing the assumed data generating mechanism under \(G\).

Definition 1 (\(A_Y\) partial isolation).

There are no causal paths from \(A_Y(G)\) to \(D_{k+1}(G)\),
\[
\forall k \in \{0, \ldots, K\},
\] (2)
where any node \(V(G)\) represented on the DAG denotes a random variable \(V\) under the treatment assignment in \(G\).

Condition (2) ensures that the treatment \(A_Y\) does not directly or indirectly affect the post-treatment event by any \(k+1\). Analogous to the causal DAG representing the data generating mechanism for the current study of Section 2 (e.g., Figure 1(a)), the causal DAG representing the mechanism for the future trial \(G\) (e.g., Figure 1(b)) relies on subject matter expertise/assumptions relative to the modified treatments \(A_Y\) and \(A_D\).

Condition (2) relates to the condition of \(A_Y\) partial isolation for time-to-event outcomes given in Stensrud et al. (2021), and therefore we will use the same terminology here.\(^3\) However, the version of \(A_Y\) partial isolation we consider here, as defined in condition (2), refers to outcomes that may not be time-to-events. \(A_Y\) partial isolation is illustrated in the causal DAGs of Figure 1(b) and Figure 2(b)–(c) (with the index \(G\) suppressed), and in subsequently presented causal DAGs under \(G\) to avoid clutter). Figure 2(c) depicts a weaker version of \(A_Y\) partial isolation by allowing the additional path \(A_D \rightarrow Z_1\) (blue), compared to that depicted in Figure 2(b). By contrast, \(A_Y\) partial isolation fails in Figure 2(d) due to the path \(A_Y \rightarrow Z_1 \rightarrow D_2\). In a causal DAG (Pearl 2000; Robins and Richardson 2010) which represents a FFRCISTG model encoding individual level effects, \(A_Y\) partial isolation (2) ensures that \(D_{k+1}^{a_Y, a_D} = D_{k+1}^{a_D}\) \forall a_Y, a_D \in \{0, 1\}\) and \(\forall a_k \in \{0, \ldots, K\}\) where \(D_{k+1}^{a_D}\) denotes the posttreatment event indicator by \(k+1\), had, possibly contrary to fact, an individual been assigned \(A_D = a_D\) under any value of \(A_Y\).\(^4\) Importantly, \(A_Y\) partial isolation is a falsifiable assumption.

\(^3\)In the mediation context, Robins and Richardson (2010, sec. 6.1) provided (using different nomenclature) examples of \(A_Y\) partial isolation (their Figure 6a) and the related conditions \(A_D\) partial isolation (their Figure 6b) and full isolation (their Figure 4) (see also Stensrud et al. (2021) and Section 6.2 in this article). Robins and Richardson showed identification of the marginal distribution of \(Y^{a_Y, a_D}\), for \(a_D \neq a_Y\) under the assumptions that \(A\) was randomly assigned and (6), which we introduce in Section 5, held. However, they showed the identifying formulas depended on whether the true causal DAG satisfied \(A_Y\) versus \(A_D\) partial isolation. In contrast to the current article Robins and Richardson (2010, sec. 6.1) did not consider interpretation and identification of conditional effects.

\(^4\)That is, under the natural value of \(A_Y\), see Richardson and Robins (2013).
For example, we can observe in the future 4-arm trial G whether
\[
\mathbb{E}[D_{k+1}(G) \mid A_Y(G) = 1, A_D(G) = a_D] = \mathbb{E}[D_{k+1}(G) \mid A_Y(G) = 0, A_D(G) = a_D],
\]
for \(k \in \{0, \ldots, K\}\) and \(a_D \in \{0, 1\}\), with failure of (3) falsifying \(A_Y\) partial isolation. The equality in (3) follows from \(A_Y\) partial isolation under either a FFRCISTG or an agnostic model (Dawid 2000; Robins and Richardson 2010; Richardson and Robins 2013; Robins, Richardson, and Shpitser 2020; Dawid 2021). Yet, under the FFRCISTG model, it is possible that (3) holds, but \(D_{k+1}^{aY,aD} = D_{k+1}^{aD}\) fails for some individuals in the population (Richardson and Robins 2013; Robins, Richardson, and Shpitser 2020). However, settings where the equality of individual-level counterfactuals holds but equality in expected counterfactuals holds must involve perfect cancelation of individual level effects. While possible, we have not encountered plausible scientific stories that justify such perfect cancelations.

The relation between \(A_Y\) partial isolation in a causal DAG representing an underlying FFRCISTG model and the equality \(D_{k+1}^{aY,aD} = D_{k+1}^{aD}\) can be more explicitly seen by minimal counterfactual labeling in a Single World Intervention Graph (SWIG), which explicitly depicts counterfactual variables Richardson and Robins (2013). For example, Figure 3(a) depicts a SWIG that is a transformation of the causal DAG in Figure 2(c), consistent with \(A_Y\) partial isolation, under an intervention that sets \(A_Y\) to \(a_Y\) and \(A_D\) to \(a_D\). Minimal labeling allows removal of the \(a_Y\) superscript on the counterfactual values under intervention of both \(Z_1\) and \(D_2\). By contrast, minimal labeling does not allow removal of this \(a_Y\) superscript in Figure 3(b) which is a corresponding transformation of the causal DAG in 2d consistent with failure of \(A_Y\) partial isolation.

Under \(A_Y\) partial isolation, the counterfactual contrast
\[
\mathbb{E}(Y^{aY=1,aD} - Y^{aY=0,aD} \mid D_{k+1}^{aD} = 0)
\]
is the average causal effect of the treatment \(A_Y\) on \(Y\) when all individuals are assigned \(A_D = a_D\) in the subset of individuals who do not experience the posttreatment event under \(A_D = a_D\), regardless of the value of \(A_Y\) they are assigned. Thus, under \(A_Y\) partial isolation this subset of individuals can be directly observed in a study that assigns \(A_Y\) and \(A_D\) as simply the subset of individuals who do not experience the posttreatment event among all those receiving \(A_D = a_D\). This is in contrast to the unobservable subset of individuals who define the principal stratum effect (1). We refer to (4) as the conditional separable effect evaluated at \(a_D \in \{0, 1\}\). We take the assumption of \(A_Y\) partial isolation as given throughout the remainder of this manuscript, unless otherwise stated. In Section 6 we will argue that this assumption is required to meaningfully define an effect of the treatment on an outcome that is only defined/of interest conditional on a posttreatment event.

In Section 6.1, we discuss two future treatments \(A_Y\) and \(A_D\) consistent with the assumption of \(A_Y\) partial isolation (2) in the cancer and quality of life example. Provided that these treatments are defined such that they can plausibly be developed and assigned in a future trial \(G\), the conditional separable effect evaluated at \(A_D = a_D\) is identified by design in that trial and can be trivially estimated by the mean difference in outcomes in the arm assigned \(A_Y = 1\) and \(A_D = a_D\) versus \(A_Y = 0\) and \(A_D = a_D\), among all those in these two arms not experiencing the post-treatment event.

5. The Modified Treatment Assumption

We now give conditions under which the conditional separable effects provide an explanation of the mechanism by which the original treatment \(A\) affects \(Y\). Consider two studies: the current study where \(A\) is randomly assigned and a future study where \(A_Y\) and \(A_D\) are jointly assigned. Following Stensrud et al. (2021, Appendix A), for two variables \(M_Y\) and \(M_D\), suppose that the following conditions hold in these two studies.

All causal paths from \(A, A_Y\) and \(A_D\) to \(Y\) and \(D_k, k \in \{0, \ldots, K\}\), are intersected by \(M_Y\) or \(M_D\),
\[
M_Y^{aY=0,aD} = M_Y^a \quad \text{for} \quad a_D \in \{0, 1\}, \quad \text{and} \quad M_D^{aY,aD=a} = M_D^a \quad \text{for} \quad a_Y \in \{0, 1\}.
\]

We refer to (5) as the modified treatment assumption. The modified treatment assumption implies that jointly assigning \(A_Y\) and \(A_D\) to the same value \(a\) leads to exactly the same values of \(Y\) and \(D_{k+1}, k \in \{0, \ldots, K\}\) as assigning \(A\) to \(a\).

Robins and Richardson (2010) introduced a decomposition assumption requiring that the treatments \(A_Y\) and \(A_D\) constitute a decomposition of \(A\) such that \(A\) exerts all its effects on \(Y\) and \(D_{k+1}\) through \(A_Y\) and \(A_D\), and the following determinism holds in the current study,
\[
A \equiv A_D \equiv A_Y.
\]

This assumption can be understood as a trivial case of (5) when we choose \(M_Y \equiv A_Y\) and \(M_D \equiv A_D\). To fix ideas, we can consider an example where (5) holds but not their
decomposition assumption. Suppose that an old chemotherapeutic treatment \((A = 1)\) exerts cytotoxic effects by alkylating many molecules in cancer cells (here, \(M_D\)). However, \(A = 1\) also has an unwanted side-effect that reduces quality of life compared to no treatment \((A = 0)\): it causes nausea by binding to certain (neurokinin 1) receptors in the brain; define this binding as \(M_Y\). A modified version of this treatment \((A_Y = 0, A_D = 1\) might be created, changing the chemical structure of the old treatment such that it no longer binds to the receptors in the brain \((M_{D_Y}^{a_Y,0,a_D=1} = M_{D_Y}^{a_Y,0})\), but still exerts the same cytotoxic effects on cancer cells and thus mortality in the same way \((M_{D_Y}^{a_Y,a_D=1} = M_{D_Y}^{a_Y,1})\). The new drug may satisfy (5) regardless of how the chemical structure of the old treatment was changed. This could involve removing a component, therefore, satisfying the treatment decomposition assumption (6). Alternatively, it could involve adding something to the old treatment that prevents the drug from binding to receptors in the brain. Interestingly, Robins and Richardson themselves (Robins and Richardson 2010) gave an example in their Section 5.2 in which rather than dividing the treatment (exogenous nicotine) into two components, they added a component to the treatment. In their example \(M_D\) was a carotid arterial nicotine receptor and \(M_Y\) was a cardiac nicotine receptor. The modified treatment, \(A_D\), added both a side group to the exogenous nicotine and a nano cage to surround the carotid receptor. In their subsequent work Robins and Richardson did not return to such examples but only considered splitting treatment into subcomponents rather than modifying treatment.

Under (5), the conditional separable effect evaluated at \(A_D = a_D\) \((4)\) is defined in the subset of the population with \(D_{k+1} \mid a_Y, a_D\) = \(D_{k+1} = 0\). To fix ideas, let \(D_{k+1}\) indicate death. Then, the modified treatment assumption (5) ensures that those who survive when assigned \(A = a_D\), would also survive when assigned \(A_D = a_D\), regardless of the value of \(A_Y\). In the current trial of Section 2, we can simply identify this subset of the population by those with \(D_{k+1} = 0\) and treatment assignment \(A = a_D, a_D \in \{0,1\}\).

Like \(A_Y\) partial isolation, assumption (5) for a choice of \(A_Y\) and \(A_D\) must be justified by working subject matter knowledge, and can be falsified in a plausible six-arm randomized experiment (denoted \(G'\)) in which individuals are randomly assigned to \(A\) (without assignment to \(A_Y\) or \(A_D\) to that is, \(A_Y\) and \(A_D\) are fixed to their null value) or joint assignment to \(A_Y\) and \(A_D\) (without assignment to \(A\)). We can for example observe in this six-arm trial whether \(E(V \mid A_Y = a, A_D = a) = E(V \mid A = a)\), for \(a \in \{0,1\}\) and \(V \in \{Y, D_{k+1}\}\), with failure of this equality falsifying (5).

Causal graphs can display a necessary condition for the modified treatment assumption (5); that is, whether \(A\), \(A_Y\) and \(A_D\) exert all their effects on \(Y\) and \(D_{k+1}\), \(k \in \{0,\ldots,K\}\) through \(M_Y\) and \(M_D\). This is illustrated in Figure 1(d), which describes the six-arm trial \(G'\) that expands the 2 arm-trial in Figure 1(a). Furthermore, the graph in Figure 1(b) can be interpreted as a transformation of the graph in Figure 1(d) that removes the node \(A\), representing the data generating mechanism under \(G'\) had we removed the two arms assigning \(A\) (i.e., a four-arm trial \(G\)). It is not necessary to include \(M_Y\) and \(M_D\) in this particular reduced graph, because they are not common causes of any variable.

If we impose the treatment decomposition assumption of Robins and Richardson (2010), we can represent the causal structure in an extended causal DAG (Robins and Richardson 2010); that is, a transformation of the original causal DAG representing \(A\) and the components \(A_Y\) and \(A_D\), where the mechanisms by which these components individually operate on outcomes are encoded (Robins and Richardson 2010; Stensrud et al. 2021). For example, Figure 1(c) is an extension of Figure 1(a) representing (6), with bold arrows representing deterministic relations. Under the decomposition assumption, Figure 1(b) can be interpreted as a \(G\)-transformation of the extended DAG in Figure 1(c), where (i) the node \(A\) and any of its causes are removed and (ii) all nodes are indexed by \(G\), with \(G\) again indexing the 4-arm trial discussed above. This \(G\)-transformation is isomorphic to a Single World Intervention Graph (SWIG) (Richardson and Robins 2013; Robins, Richardson, and Shpitser 2020), with interventions on \(A_Y\) and \(A_D\).

6. \(A_Y\) Partial Isolation and Meaningful Effects of \(A\) on \(Y\)

By explicitly considering modified treatments such that \(A_Y\) partial isolation (2) and the modified treatment assumption hold, the investigator is forced to articulate what she means by an “effect of \(A\) on \(Y\) not through \(D_{K+1}\)” under assumptions that are falsifiable in a future experiment. This thought process requires the investigator to be explicit about her notion of a causal mechanism, and allows the consideration of a well-defined causal effect: in particular, under these assumptions, the conditional separable effects quantify mechanisms by which \(A\) affects \(Y\) that can be entirely separated from mechanisms by which \(A\) affects \(D_{K+1}\), \(k \in \{0,\ldots,K\}\). We discuss this further in Section 6.2.

If the investigator is unable to express a convincing story about modified treatments satisfying (2) and the modified treatment assumption, then the relevance of an effect of \(A\) on \(Y\) outside of its effect on the posttreatment event \(D_{K+1}\) is ambiguous: the investigator has failed to give a plausible scientific argument as to how effects of \(A\) on \(Y\) can be disentangled from effects of \(A\) on \(D_{K+1}\). Yet, even if the investigator cannot define plausible modified treatments satisfying (2) and the modified treatment assumption at this moment in time, these assumptions may be justified in the future: modified treatments satisfying these assumptions might be revealed when more subject-matter knowledge becomes available. However, until a plausible story can be articulated such that (2) and the modified treatment assumption are satisfied, the practical relevance of considering any effect of \(A\) on \(Y\) outside of its effect on \(D_{K+1}\)—including the conventional principal stratum effect—is unclear. In this case, the investigators must accept that they do not understand how \(A\) exerts such effects and have no way to assess whether such effects are operating in the data without reliance on assumptions that are impossible to ever challenge in real-life experiments. In turn, without any ideas about such modified treatments, it is unclear whether any feasible interventions can avoid or leverage effects of \(A\) on \(Y\) outside of its effect on \(D_{K+1}\) (Robins and Richardson 2010).
6.1. Example: Cancer Treatment and Quality of Life

Returning to our cancer treatment and quality of life example, suppose that Figure 2(a) represents data generating assumptions on a trial that assigns treatment at baseline (A = 1 is new chemotherapy, A = 0 is standard chemotherapy) with Y a quality of life measure at end of follow-up and Dk+1 an indicator of death by time k + 1. Suppose that there exist two modified treatments \( A_Y \) and \( A_D \) satisfying the following assumptions, which are also illustrated in Figure 2(c): \( A_D \) exerts effects on mortality \( D_1, D_2 \), for example, by destroying or reducing the growth of cancer cells and thereby preventing cancer progression \((Z_1)\). By preventing cancer progression, the \( A_D \) component may also reduce other health problems, for example, due to metastases, which affect quality of life \( Y \) (the path \( A_D \rightarrow Z_1 \rightarrow Y \) in Figure 2(c)). The other component \( A_Y \) does not exert effects on mortality, because it has little to no activity against the cancer but may have side effects that adversely affect quality of life; for example it may interfere with the replication of epithelial mucosal cells, resulting in diarrhea and oral ulcers. Alternatively, this component may possibly have beneficial effects (say due to a decrease in diarrhea and oral ulcers).

In this setting, the conditional separable effect evaluated at \( a_D = 1 \) quantifies the treatment effect on quality of life outside of its effect on disease progression. Specifically, this quantifies the effect of assignment to a current chemotherapy (e.g., \( a = a_Y = a_D = 1 \)) versus a modified (hypothetical) therapy that contains the component of the current therapy that reduces mortality and disease progression \((a_D = 1, a_Y = 0)\). An improvement of quality of life under the modified therapy suggests that the current chemotherapy \((a = 1)\) contains a component that it would be desirable to eliminate.

\( A_Y \) partial isolation would fail to hold in our example if \( A_Y \) exerts effects on a common cause of \( Y \) and \( D_{k+1}, k = 0, \ldots, K+1 \), as illustrated in Figure 2(d) by the path \( A_Y \rightarrow Z_1 \rightarrow D_2 \). For example, suppose now that \( Z_k \) for \( k < K \) denotes quality of life at times earlier than \( Y \), where \( Y \) indicates quality of life at the end of follow-up \( K \). If the \( A_Y \) component exerts effects on quality of life only after a minimal latent period that extends beyond the study period \((\text{Robins 2008})\), then \( A_Y \) partial isolation may still be justified (e.g., the arrow from \( A_Y \) into \( Z_1 \) in Figure 2(d) can be removed). Alternatively, if quality of life only exerts effects on mortality after a minimal latent period that extends beyond the study period, then \( A_Y \) partial isolation may be justified (e.g., the arrow from \( Z_1 \) into \( D_2 \) can be removed).

6.2. When the Conditional Separable Effect is the Direct Effect of A

While the conditional separable effects are well-defined under \( A_Y \) partial isolation, this condition allows additional causal paths from \( A \) to \( Y \) that are not intersected by \( D_{k+1}, k = 0, \ldots, K \), as illustrated in Figure 2(c) by the path \( A_D \rightarrow Z_1 \rightarrow Y \). In this sense, under \( A_Y \) partial isolation, a conditional separable effect captures a direct effect, but not all effects of \( A \) on \( Y \) outside of \( D_{k+1} \) relative to the posttreatment event. However, suppose that, in a causal DAG representing the assumed data generating mechanism in the four-arm trial \( G \) the following condition holds.

**Definition 2 (A_D partial isolation).**

The only causal paths from \( A_D(G) \) to \( Y(G) \) are directed paths intersected by \( D_k(G), k \in \{1, \ldots, K\} \).

When both \( A_D \) partial isolation (7) and \( A_Y \) partial isolation (2) simultaneously hold, we say there is full isolation (Stenrud et al. 2021). Under full isolation, the conditional separable effects capture all causal paths from \( A \) to \( Y \) not intersected by \( D_{k+1}, k = 0, \ldots, K \). Full isolation is represented in Figures 1(b) and 2(b).5

Returning to our running cancer treatment and quality of life example represented by Figure 2(c), full isolation would hold under the stronger assumption that cancer progression \( Z_k, k = 0, \ldots, K+1 \) does not affect quality of life \( Y \), allowing removal of the arrow from \( Z_1 \) to \( Y \). This assumption seems to be implausible for many cancer treatments: by preventing cancer progression, the treatment will not only reduce mortality, but also reduce other effects of progression, such as pain related to tumor growth.

7. Identifiability Conditions

If we had data from a four-arm trial in which \( A_Y \) and \( A_D \) were randomly assigned and censoring were absent, conditions required to identify \( \mathbb{E}(Y^{a_Y,a_D} | D_{k+1}^{a_Y,a_D} = 0) \) for \( a_Y, a_D \in \{0,1\} \) hold by design. Consequently, these conditions also identify the conditional separable effects under the assumption of \( A_Y \) partial isolation using that

\[
\mathbb{E}(Y^{a_Y,a_D} | D_{k+1}^{a_Y,a_D} = 0) = \mathbb{E}(Y^{a_Y,a_D} | D_{k+1}^{a_Y,a_D} = 0).
\] (8)

However, in the two-arm trial in which only the original treatment \( A \) is randomly assigned, we are not guaranteed identification of \( \mathbb{E}(Y^{a_Y,a_D} | D_{k+1}^{a_Y,a_D} = 0) \) when \( a_Y \neq a_D \) in this trial even when censoring is absent.

We now consider a set of conditions, beyond the conventional exchangeability, consistency and positivity conditions that hold by design in the two-arm trial of Section 2 (reviewed in supplementary material A). Under the modified treatment assumption, these conditions are sufficient to identify \( \mathbb{E}(Y^{a_Y,a_D} | D_{k+1}^{a_Y,a_D} = 0) \) when \( a_Y \neq a_D \) using only data from this existing trial. By (8), this allows identification of the conditional separable effects.

**Assumption (Positivity).**

\[
f_{I_k,D_{k+1}}(i_k, 0) > 0 \implies f_{I_k,D_{k+1} | A}(i_k, 0 | a) > 0, \text{ for all } k \in \{0, \ldots, K\}, a \in \{0,1\}.
\] (9)

Assumption (9) states that for any possibly observed level of the time-varying covariate history among those surviving

---

5 Figure 2(b) illustrates that condition (7) does not imply that \( Y \perp \perp A_D | D_{k+1}, A_Y \), for example, due to the collider path \( A_D \rightarrow D_2 \leftarrow Z_1 \rightarrow Y \).
through each follow-up time, this covariate history level and survival can be observed among individuals with $A = 1$ and individuals with $A = 0$. Assumption (9) does not hold by design in a randomized experiment, but it can be assessed in the observed data.

**Assumption (Discernible component conditions).**

\[
Y(G) \perp\!\!\!\!\!\!\!\perp A_D(G) \mid A_Y(G), D_{K+1}(G) = 0, L_K(G),
\]

(10)

\[
D_{K+1}(G) \perp\!\!\!\!\!\!\!\perp A_Y(G) \mid A_D(G), D_k(G) = 0, L_k(G),
\]

(11)

\[
L_{k+1}(G) \perp\!\!\!\!\!\!\!\perp A_Y(G) \mid A_D(G), D_{k+1}(G) = 0, L_k(G),
\]

(12)

for all $k \in [0, \ldots, K]$. The discernible component conditions (10)–(12) can be assessed in G-transformation graphs as discussed in Section 5, using d-separation rules (Liver Flaf 2000; Robins and Richardson 2010).6 These conditions can be interpreted as probabilistic independencies in the trial G.7

For example, Figure 4 depicts the assumption that there exist measured (e.g., $L_1$) and unmeasured (e.g., $L_{U_1}$ or $L_{U_D}$) common causes of $Y$ and $D_2$ in G. The discernible component conditions hold in Figure 4(a)–(c), but (10) is violated in Figure 4(d)–(e) and (12) is violated in Figure 4(f). These examples illustrate that the discernible component conditions (10)–(12) are sufficient (see the lemma in supplementary material B) but not necessary for $A_Y$ partial isolation. For example, $A_Y$ partial isolation holds in Figure 4(d)–(e) even though discernible component conditions are violated. $A_Y$ partial isolation fails in Figure 4(f).

**Theorem 1.** Under consistency, conditional exchangeability and assumptions (9)–(12), the conditional counterfactual mean $E(Y_{a_D, a_Y}|D_{K+1} = 0)$ is identified by

\[
\sum_{L_K} E(Y \mid D_{K+1} = 0, L_K = l_K, A = a_y) \\
\frac{f_{L_1,D_{K+1}|L_0,A}(l_1,0 \mid l_0,D_D) f_{L_0}(l_0)}{P(D_{K+1} = 0 \mid A = a_D)}.
\]

(13)

See supplementary material A for a proof that also covers settings where individuals can be lost to follow-up.8 We say that (13) is the g-formula for $E(Y_{a_D, a_Y}|D_{K+1} = 0)$ (Robins 1986).

Importantly, $A_Y$ partial isolation clarifies when a contrast of conditional counterfactual means,

\[
E(Y_{a_Y=1,a_D}|D_{K+1} = 0) \text{ versus } E(Y_{a_Y=0,a_D}|D_{K+1} = 0),
\]

(14)

can be interpreted as a conditional causal effect: even when treatment $A_Y$ is temporally ordered before $D_{K+1}$, $A_Y$ partial isolation allows us to topologically order the component $A_Y$ after $D_{K+1}$. That is, our estimands are isomorphic to estimands defined by interventions where we first assign a treatment $A_D$ and then, after $D_{K+1}$ occurs, we subsequently assign $A_Y$. This can be seen on a SWIG with minimal labeling (Richardson and Robins 2013) consistent with $A_Y$ partial isolation, as in Figure 3(a). In supplementary material A we show that $A_Y$ partial isolation is actually not in general needed to identify $E(Y_{a_Y,a_D}|D_{K+1} = 0)$ but is needed to ensure that (14) is an average over individual level causal effects under an (individual level) FFRCISTG model.9

### 7.1. Related Works on Identification of Path Specific Effects

There is an intimate link between identification results for our estimands and path-specific effects, as recently discussed in Robins, Richardson, and Shpitser (2020): when there is no so-called recanting witness (Avni, Shpitser, and Pearl 2005; Shpitser 2013; see also Stensrud et al. 2021, sec. 6.5), the (cross-world) counterfactuals defining a path-specific effect are equal to (single-world) counterfactuals defined by interventions on nodes in an extended causal graph, like Figure 1(c). Thus, the identification formulas for effects defined by interventions in an extended causal graph are equal to identification formulas derived for certain path-specific effects (Robins, Richardson, and Shpitser 2020). Malinsky, Shpitser, and Richardson (2019) recently derived a complete algorithm for identification of conditional path-specific distributions, which is inspired by the general identification theory of Shpitser (2013).10 However, while the identification formulas for these estimands can be equal, the effects being identified are different: they refer to interventions on different variables (see Robins, Richardson, and Shpitser 2020 for a detailed discussion). Unlike our current results, the previous works did not consider conditions under which a contrast of conditional counterfactual outcome parameters can be interpreted as a conditional causal effect (i.e., a contrast of counterfactual outcome means in the same set of individuals).

### 7.2. New Identification Results for Principal Stratum Effects

For the principal stratum effect (1) to target the same subpopulation as the conditional separable effects (4) under both $a_D = 0$ and $a_D = 1$, we must make the additional strong assumption that $D_{k+1} = D_k^m = 0$ for $k \in [0, \ldots, K]$. For example, under the assumption that lack of arrows in a causal graph means no individual level causal effect, this requires that the arrow from

8More specifically, we can use D-separation to read off stronger versions of (10)–(12), where $D_j(G) \in \{k+1, K\}$ is not instantiated, that is, we replace $D_j(G) = 0$ by $D_j(G)$.

9Using theory of extended conditional independence in an agnostic causal model, the discernible component conditions would also be valid for nonrandom interventions on decision variables (Constantinou and Dawid 2017), as pointed out by an anonymous reviewer. This corresponds to nonrandom interventions in a population SWIG model; Richardson and Robins (2013, sec. 9) describe how investigators who do not want to assume the existence of counterfactual outcomes, still can construct a graph isomorphic to a SWIG.

8In supplementary material A, the proof for the identification formula of $E(Y_{a_Y,a_D}|D_{K+1} = 0)$ is given in a more general setting where $A_Y$ partial isolation is not required to hold. However, $A_Y$ partial isolation is a necessary condition for the conditional separable effects to be well-defined.

9Under a population level SWIG model (Robins and Richardson 2010; Richardson and Robins 2013) and an agnostic model (Richardson and Dawid 2021), the contrast (14) is not guaranteed to be an average over individual level effects, but it still quantifies the contrast of counterfactual outcomes under treatments $A_Y = 1$ versus $A_Y = 0$ when $A_D$ is fixed to $a_D$, and it still holds that $E(D_{K+1}|a_Y = 1, a_D) = E(D_{K+1}|a_Y = 0, a_D)$. Thus, (14) captures the contrast in average outcomes under a regime that assigns $A_Y = 0, A_D = a$ versus a regime that assigns $A_Y = 1, A_D = a$.

10These results cover conventional estimands, such as pure (natural) direct and indirect effects, as a special case.
A into $D_1$ in Figure 1(a) or the arrows from $A$ into $D_1$ and $D_2$ in Figure 2(a) should be removed. In our cancer and quality of life example, this assumption requires survival to be identical under the new ($a = 1$) and standard ($a = 0$) chemotherapy. This underscores that the conditional separable effects and the principal stratum effect (1)—specifically the survivor average causal effect (SACE) in this example—are substantially different estimands; these estimands will only be equivalent in the restrictive setting where $A$ exerts no effect on mortality.

Suppose instead that we make the weaker assumption that the effect of $A$ on $D_{k+1}$ is monotone: without loss of generality, this is the assumption that $D_{k+1}^a = D_{k+1}^0$ in all individuals for all $k$. Then, the following proposition gives a relation between principal stratum effects and separable effects.

**Proposition 1.** Suppose that $D_{k+1}^a = D_{k+1}^0$, and that the modified treatment assumption (5) and full isolation holds. Then, the principal stratum effect (1) is equivalent to the conditional separable effect under $aD = 0$,

$$
E(Y_{a=1} - Y_{a=0} \mid D_{k+1}^a = D_{k+1}^0 = 0) = E(Y_{aY=1,aD=0} - Y_{aY=0,aD=0} \mid D_{k+1}^a = D_{k+1}^0 = 0).
$$

**Proof.**

$$
E(Y_{a=1} - Y_{a=0} \mid D_{k+1}^a = D_{k+1}^0 = 0) = E(Y_{aY=1,aD=0} - Y_{aY=0,aD=0} \mid D_{k+1}^a = D_{k+1}^0 = 0)
$$

by monotonicity

$$
= E(Y_{aY=1,aD=0} - Y_{aY=0,aD=0} \mid D_{k+1}^a = D_{k+1}^0 = 0) \quad \text{by (7)}
$$

It follows directly from Proposition 1 that full isolation, the modified treatment assumption, and conditions (9)–(12) are sufficient to identify the principal stratum effect (1) under monotonicity in the two-arm trial. Thus, our identification results supplement the few suggested identifiability assumptions that are sufficient for identification of principal stratum effects, such as the survivor average causal effect, in the presence of measured, possibly time-varying, common causes of the event.
of interest and the posttreatment conditioning event Tchetgen Tchetgen (2014). Furthermore, we rely on assumptions that can be falsified (rejected) in a future randomized experiment.

8. Estimation

Let $v_{ay,ad}$ denote the g-formula (13). Here we consider various estimators for this parameter in the absence of censoring. Extensions to allow censoring are given in supplementary material D.

8.1. Outcome Regression Estimator

A simple outcome regression estimator $\hat{v}_{or,ay,ad}$ of $v_{ay,ad}$ is the solution to the estimating equation $\sum_{j=1}^{n} U_{or,i}(v_{ay,ad}, \hat{\theta}) = 0$ with respect to $v_{ay,ad}$, with

$$U_{or,i}(v_{ay,ad}, \hat{\theta}) = (A_i = a_y)(1 - D_{K+1,i}) \times \left( \mathbb{E}(Y | D_{K+1} = 0, A = a_Y, \hat{L}_{K,i}; \hat{\theta}) - v_{ay,ad} \right),$$

where $\mathbb{E}(Y | D_{K+1} = 0, A = a_Y, \hat{L}_{K}; \hat{\theta})$ is a parametric model for $\mathbb{E}(Y | D_{K+1} = 0, \hat{L}_{K}; A = a_Y)$ indexed by the parameter $\theta$ and assume $\hat{\theta}$ is its MLE. The estimating equation $U_{or,i}(v_{ay,ad}, \hat{\theta})$ has mean zero and the estimator $\hat{v}_{or,ay,ad}$ is consistent provided that this model is correctly specified.

8.2. Weighted Estimator

Alternatively, define the weighted estimator $\hat{v}_{pw,ay,ad}$ of $v_{ay,ad}$ as the solution to the estimating equation $\sum_{j=1}^{n} U_{pw,i}(v_{ay,ad}, \hat{\alpha}) = 0$ with respect to $v_{ay,ad}$, with

$$U_{pw,i}(v_{ay,ad}, \hat{\alpha}) = I(A_i = a_y)(1 - D_{K+1,i}) \hat{W}_i(a_y, a_D; \hat{\alpha}) \left( Y_i - v_{ay,ad} \right),$$

such that

$$\hat{W}(a_y, a_D; \hat{\alpha}) = \frac{f_{\hat{L}_{j,i}}[\hat{\theta}]}{f_{\hat{L}_{j,i}}[\theta]} a_Y I(\hat{L}_{j,i} = 1, 0 | L_0, a_D; \hat{\alpha})$$

is an estimator of

$$W(a_y, a_D) = \frac{f_{\hat{L}_{j,i}}[\hat{\theta}]}{f_{\hat{L}_{j,i}}[\theta]} a_Y I(\hat{L}_{j,i} = 1, 0 | L_0, a_D),$$

where $f_{\hat{L}_{j,i}}[\hat{\theta}](l_0, 0 | l_0, a)$ is a parametric model for $f_{\hat{L}_{j,i}}[\theta](l_0, 0 | l_0, a)$ indexed by the parameter $\alpha$ and assume $\hat{\alpha}$ is its MLE. If this model is correctly specified, then $\hat{v}_{pw,ay,ad}$ is a consistent estimator for $v_{ay,ad}$, which follows from an alternative representation of the g-formula that is derived in supplementary material A.1.

In practice, parameterizing the terms in $W(a_y, a_D)$ requires care when $\hat{L}_K$ is high-dimensional. However, we can re-express $\hat{W}(\cdot)$ as a product of the terms

$$\hat{W}_D(a_y, a_D; \hat{\alpha}) = \frac{\prod_{j=0}^{K} \text{Pr}(D_{j+1} = 0 | \hat{L}_{j}, D = 0, A = a_D; \hat{\alpha})}{\prod_{j=0}^{K} \text{Pr}(D_{j+1} = 0 | \hat{L}_{j}, D = 0, A = a_D; \hat{\alpha})},$$

$$\hat{W}_L(a_y, a_D; \hat{\alpha}) = \frac{\prod_{j=0}^{K-1} f_{\hat{L}_{j+1}}[\hat{L}_{j+1} | \hat{L}_{j}, a_Y; \hat{\alpha}] \hat{L}_{j+1} | 0, \hat{L}_{j}, a_Y; \hat{\alpha})}{\prod_{j=0}^{K-1} f_{\hat{L}_{j+1}}[\hat{L}_{j+1} | \hat{L}_{j}, a_Y; \hat{\alpha}] \hat{L}_{j+1} | 0, \hat{L}_{j}, a_Y; \hat{\alpha})},$$

where $\text{Pr}(D_{j+1} = 0 | D_j = 0, \hat{L}_j, A = a_D)$ is a pooled over time model for $\text{Pr}(D_{j+1} = 0 | D_j = 0, \hat{L}_j, A = a_D)$ indexed by the parameter $\alpha_D$, and $\hat{\alpha}_D$ its MLE. The parameter $\alpha_D$ analogously indexes a pooled over time model for the conditional density $f_{L_{j+1}}[\hat{L}_{j+1} | \hat{L}_j, A](l_{j+1} | 0, \hat{L}_j, a)$, with $\hat{\alpha}_L$ its MLE. Off the shelf software can be used to estimate $\text{Pr}(D_{j+1} = 0 | D_j = 0, \hat{L}_j, A = a_D)$. More complex computation (and strong parametric assumptions) may be required to consistently estimate $f_{L_{j+1}}[\hat{L}_{j+1} | \hat{L}_j, A](l_{j+1} | 0, \hat{L}_j, a)$ when $L_j$ is high-dimensional. In this case, we can alternatively fit a model directly for the likelihood ratio

$$\frac{f_{L_{j+1}}[\hat{L}_{j+1} | \hat{L}_j, A](l_{j+1} | 0, L_j, a)}{f_{L_{j+1}}[\hat{L}_{j+1} | \hat{L}_j, A](l_{j+1} | 0, \hat{L}_j, a_Y)};$$

for $j \in \{0, \ldots, K\}$, for example, a proportional likelihood ratio model as a more parsimonious function of $L_j$ and possibly $j$ which avoids distributional assumptions on $L_{j+1}$ (Luo and Tsai 2012).

8.3. Doubly Robust Estimator

We can alternatively consider a doubly robust estimator derived from the nonparametric influence function in supplementary material C (Robins, Rotnitzky, and Zhao 1994; Van der Vaart 2000; Van der Laan and Robins 2003). This estimator $\hat{v}_{dr,ay,ad}$ is the solution to the estimating equation $\sum_{j=1}^{n} U_{dr,i}(v_{ay,ad}, \hat{\alpha}, \hat{\theta}) = 0$ with respect to $v_{ay,ad}$, with

$$U_{dr,i}(v_{ay,ad}, \hat{\alpha}, \hat{\theta}) = \left( I(A_i = a_y) - \frac{P(A_i = a_y)}{P(A_i = a_y)} (1 - D_{K+1,i}) \right) \times \left( \frac{I(A_i = a_y)}{P(A_i = a_y)} \mathbb{E}(Y | D_{K+1} = 0, A = a_Y, \hat{L}_{K,i}; \hat{\theta}) + \frac{I(A_i = a_y)}{P(A_i = a_y)} \frac{f_{\hat{L}_{i}}[\hat{L}_{j+1} | L_0, a_D; \hat{\alpha}]}{f_{\hat{L}_{i}}[\hat{L}_{j+1} | L_0, a_Y; \hat{\alpha}]} \right) \times \left( Y_i - \mathbb{E}(Y | D_{K+1} = 0, A = a_Y, \hat{L}_K; \hat{\theta}) \right) \times \left( \frac{1}{\mathbb{E}(1 - D_{K+1} | A = a_D) - v_{ay,ad}} \right),$$

with $\mathbb{E}(1 - D_{K+1} | A = a_D)$ simply a sample mean. In supplementary material C we show that $\hat{v}_{dr,ay,ad}$ is doubly robust; that is, it is consistent if either the models for $f_{\hat{L}_{i}}[\hat{L}_{j+1} | L_0, a_D; \hat{\alpha}]$ or $\mathbb{E}(Y | D_{K+1} = 0, \hat{L}_K, A = a_D)$ are correctly specified, but not necessarily both. Analogous to the weighted estimator $\hat{v}_{pw,ay,ad}$, consistency of $\hat{v}_{dr,ay,ad}$ may also be achieved by correctly specifying a model for the likelihood ratio in place of $f_{\hat{L}_{i}}[\hat{L}_{j+1} | L_0, a_D; \hat{\alpha}]$, as described in the previous section. Note that the estimating equation for $\hat{v}_{dr,ay,ad}$ can serve as the basis for constructing estimators that use machine learning to estimate nuisance parameters in place of parametric models.

9. Data Example: The Southwest Oncology Group Trial

As an illustration, we analyzed data from the Southwest Oncology Group (SWOG) Trial (Petrylak et al. 2004) that randomly
assigned men with refractory prostate cancer to either of two chemotherapies, Docetaxel and Estramustine (DE) or Mitoxantrone and Prednisone (MP). Our dataset, which included 487 patients aged 47–88 years, has previously been used to compare outcomes under DE versus MP on health related quality of life in the principal stratum of always survivors (Ding et al. 2011; Wang, Zhou, and Richardson 2017; Yang and Ding 2018), that is, to estimate the survivor average causal effect. Yet, the practical relevance of the survivor average causal effect is ambiguous, as we discussed in Section 6.

The conditional separable effects quantify notions of causal mechanisms and can be used to conceive improved treatments. Thus, our target of inference was a conditional separable effect of DE versus MP on the outcome \( Y \), defined as the change in quality of life between baseline and one year of follow-up, similar to the previous reports evaluating the SACE (Ding et al. 2011; Wang, Zhou, and Richardson 2017; Yang and Ding 2018). We considered a modification of the original treatments DE (\( A = 1 \)) and MP (\( A = 0 \)), where \( A_D = 1 \) indicates receiving only the Docetaxel component of DE and \( A_D = 0 \) indicates receiving only Mitoxantrone component of MP. Both Docetaxel and Mitoxantrone are chemotherapeutic agents that can reduce proliferation of cancer cells, and, potentially, progression of the disease (\( Z_k \)). Further, let \( A_Y = 1 \) indicate receiving only Estramustine component of DE and \( A_Y = 0 \) indicate receiving only the Prednisone component of MP. Assume that \( A_Y \) partial isolation is satisfied. Clearly, this condition is not guaranteed by design, but may be plausible for these choices of \( A_Y \) and \( A_D \). Specifically, neither Estramustine, which consists of estrogen and chemotherapeutic medication, nor Prednisone, a steroid, are given to reduce mortality. However, these components can potentially give pain relief (palliation) (Petrylak 2005), but may also have side-effects, including nausea, fatigue and vomiting (Petrylak et al. 2004; Petrylak 2005; Autio, Scher, and Morris 2012), which in turn can affect quality of life. Therefore, it is not clear which of these components are most effective at improving quality of life.

The causal structure of this example is represented in Figure 2(c), satisfying \( A_Y \) partial isolation, where the causal path \( A_D \rightarrow Z_1 \rightarrow Y \) illustrates that \( A_D \), the chemotherapeutic components Docetaxel or Mitoxantrone, can affect quality of life but only through effects on the cancer progression indicator \( Z_k \).

After \( K + 1 = 12 \) months of follow-up, 0.79 (95% CI: [0.73, 0.84]) survived in the DE arm compared to 0.72 (95% CI: [0.66, 0.78]) in the MP arm, resulting in an estimated additive causal effect of 0.07 (95% CI: [−0.01, 0.15]) of treatment \( A \) on 12-month survival (\( D_{12} \)). Of the 368 survivors at 12 months, 152 had a measure of the outcome quality of life measured (i.e., were uncensored). All analyses were adjusted for the measured baseline covariates: age, race, anticipated of prognosis, bone pain and performance status (\( L_0 \)) as well as an indicator of disease progression by \( k \) (\( L_k \)). These point estimates are in line with a possible improvement in mortality under DE compared to MP. By our hypotheses above, this suggests that the effect of a modified version of DE that replaces the Estramustine component with Prednisone (i.e., the joint treatment \( A_D = 1, A_Y = 0 \)) is of interest with respect to quality of life.

The estimated mean of \( Y \) among survivors was −4.4 (95% CI: [−8.6, −0.5]) units in the DE arm (\( a_D = a_Y = 1 \)) and −9.1 (95% CI: [−14.0, −3.9]) units in the ME arm (\( a_D = a_Y = 0 \)) (Table 1). However, we cannot assign a causal interpretation to a contrast of these estimates, as discussed in Section 1. Therefore, we estimated the conditional separable effect for \( a_D = 1 \) (i.e., the comparison of outcomes under Estramustine versus Prednisone both given with Docetaxel) using the regression estimator \( \hat{v}_{or,ay,aD} \), the weighted estimator \( \hat{v}_{ipw,ay,aD} \) based on the weights \( \hat{W}(a_Y, a_D; \hat{a}) \) as well as the doubly-robust estimator \( \hat{v}_{dr,ay,aD} \). We assumed the following parametric models for the nuisance parameter \( f_L, \pi_{K+1|L_0,A}(L_k | 0, a_D) \),

\[
\hat{\text{logit}}[\hat{\text{Pr}}(D_{12} = 1 | A = 0, L_0, L_{11}; \alpha_{D_1})] = \alpha_{D,0} + \alpha_{D,1} L_0 + \alpha_{D,2} L_{11}, \\
\hat{\text{logit}}[\hat{\text{Pr}}(D_{12} = 1 | A = 1, L_0, L_{11}; \alpha_{D_1})] = \alpha_{D,3} + \alpha_{D,4} L_0 + \alpha_{D,5} L_{11}, \\
\hat{\text{logit}}[\hat{\text{Pr}}(L_{11} = 1 | A, L_0; \alpha_{L_1})] = \alpha_{L,0} + \alpha_{L,1} A + \alpha_{L,2,0} L_0,
\]

and the following model for \( \hat{\text{E}}(Y | D_{12} = 0, A = a_Y, L_{11} = \tilde{L}_{11}) \),

\[
\hat{\text{E}}(Y | D_{K+1} = 0, A = a, \tilde{L}_K = \tilde{L}_K; \theta) = \theta_0 + \theta_1 A + \theta_2 L_0 + + \theta_3 L_{11}.
\]

Under the hypothetical treatment that contains the component of DE that affects mortality (Docetaxel, \( a_D = 1 \)), but contains the component of MP that potentially affects quality of life outside of mortality (Prednisone, \( a_Y = 0 \)), the estimated mean outcome after 12 months was −6.1 (95% CI [−11.8, −0.7]) using the simple regression estimator \( \hat{v}_{or,ay,aD} \) (Table 1). The estimated additive conditional separable effect for \( a_D = 1 \) was equal to −1.7 (95% CI [−5.1, 8.1]).

Similarly, using the simple weighted estimator \( \hat{v}_{ipw,ay,aD} \) the estimated mean outcome after 12 months was −5.5 (95% CI [−9.7, −0.4]) (Table 1), and the estimated additive conditional separable effect for \( a_D = 1 \) was equal to −1.1 (95% CI [−5.6, 6.6]). We also implemented the doubly robust estimator \( \hat{v}_{dr,ay,aD} \), which gave an estimated outcome mean of −5.6 (95% CI [−11.5, −0.2]) (Table 1), and an estimated conditional separable effect for \( a_D = 1 \) equal to −1.7 (95% CI [−5.1, 8.1]). Thus, the estimates from \( \hat{v}_{or,ay,aD}, \hat{v}_{ipw,ay,aD} \) and \( \hat{v}_{dr,ay,aD} \) are similar. The confidence intervals are wide, and there is no clear evidence that the modified drug with Docetaxel and Prednisone would lead to improved quality of life compared to DE in this study. These conclusions, however, rely on the assumption of \( A_Y \) partial isolation and the identifiability conditions in Section 7. In particular, the presence of unmeasured common causes of mortality and quality of life would violate these assumptions.
As discussed in Section 7.2, under the additional assumption of monotonicity and \( A_D \) partial isolation, our results also allow identification and estimation of the SACE. However, we have already argued that the practical relevance of the SACE is ambiguous. Under the additional assumption of a monotonic effect of \( A_D \) on mortality, that is, \( Y_{12}^{a_D=1} \leq Y_{12}^{a_D=0} \), the SACE is identified by the same functional as the conditional separable effect where \( A_D \) is fixed to 0,

\[
\mathbb{E}(Y_{12}^{a_Y=0,a_D=0} \mid X_{12}^{a_D=0}) = 0
\]

that is, the conditional separable effect among those who would survive under Mitoxantrone, regardless of whether they received Prenisone or Estramustine. In contrast, we have argued that the separable effect among those who would survive under Docetaxel (where \( a_D \) is fixed to 1) is of primary interest. Furthermore, we do not believe that monotonicity holds in our example, that is, that Mitoxantrone does not improve survival in any single individual: different individuals can experience different effects of Docetaxel and Mitoxantrone on survival, for example, due to differences in the cancer cells (subtypes) across individuals.

The only form of censoring in this dataset was due to missing values of \( Y \). In supplementary material D, we describe how censoring was adjusted for in the analysis.

10. Discussion

This work presents new estimands, the conditional separable effects, in settings where the outcome of interest is only defined conditional on a posttreatment event status. To define the conditional separable effects, we rely on the assumption of \( A_Y \) partial isolation and the existence of two modified treatments that exert effects through distinct causal pathways (Robins and Richardson 2010; Didelez 2018; Stensrud et al. 2021, 2022).

Principal stratum effects have often been used to quantify causal effects on an outcome that is only of interest conditional on a posttreatment variable, such as survival (Robins 1986; Frangakis and Rubin 2002) or having an infection (Hudgens, Hoering, and Self 2003), but the practical relevance of these estimands has been questioned, because (i) they are defined in an unknown subgroup that may not even exist (Robins, Rotnitzky, and Vansteelandt 2007; Joffe 2011; Young et al. 2018), and (ii) estimating the principal stratum effects requires unfalsifiable assumptions that cannot be verified even in any randomized experiment (Robins and Richardson 2010).

Unlike principal stratum effects, the conditional separable effects can be identified without relying on unfalsifiable independence assumptions or hypothetical interventions to prevent death; the conditional separable effects can be identified in a FFRCISTG model (Robins 1986; Robins and Richardson 2010; Richardson and Robins 2013). As such, the identifiability conditions can be scrutinized in a future experiment in which the treatment components \( A_Y \) and \( A_D \) are randomly assigned. Unlike principal stratum effects, the conditional separable effects are not restricted to an unknown subgroup of unknown size, but the two versions of the conditional separable effects (evaluated under either \( a_D = 0 \) or \( a_D = 1 \)) are defined in the (observable) subsets of individuals who would have a particular value of a posttreatment variable under assignment \( A = 0 \) and \( A = 1 \), respectively.

Thus, the conditional separable effects overcome key concerns with principal stratum estimands, such as the survivor average causal effects. Yet the definition of the conditional separable effects hinges on \( A_Y \) partial isolation, which relies on subject matter justification in any given practical setting. However, this is not a limitation of the estimand itself: unless we conceive plausible modified treatments meeting \( A_Y \) partial isolation, it is not clear how to practically disentangle treatment effects on the posttreatment variable (e.g., death) and treatment effects on the outcome of interest. Furthermore, our results help researchers clarify their thinking about future, potentially improved treatments. The study of these future treatments is of scientific and public health interest, even if \( A_Y \) partial isolation is falsified in a future experiment evaluating these treatments.

Supplementary Materials

The supplementary material includes proof of the identification formula (A), a proof showing that the dismissible component conditions imply \( A_Y \) partial isolation (B), derivations of the non-parametric influence function and doubly robust estimators (C), and a consideration of censoring (D).

ORCID

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

References

Avin, C., Shpitser, I., and Pearl, J. (2005), “Identifiability of Path-Specific Effects,” in Proceedings of International Joint Conference on Artificial Intelligence. [2678]

Bornkamp, B., Rufibach, K., Lin, J., Liu, Y., Mehrrotra, D. V., Roychoudhury, S., Schmidli, H., Shentu, Y., and Wolbers, M. (2021), “Principal Stratum Strategy: Potential Role in Drug Development,” Pharmaceutical Statistics, 20, 737–751. [2672]

Constantinou, P., and Philip Dawid, A. (2017), “Extended Conditional Independence and Applications in Causal Inference,” The Annals of Statistics, 45, 2618–2653. [2678]

Dawid, P. (2021), “Decision-Theoretic Foundations for Statistical Causality,” Journal of Causal Inference, 9, 39–77. [2675, 2678]

Dawid, P., and Didelez, V. (2012), “Imagine a Can Opener”—The Magic of Principal Stratum Analysis,” The International Journal of Biostatistics, 8, 19. [2671, 2673]

Dawid, P. A. (2000), “Causal Inference Without Counterfactuals,” Journal of the American statistical Association, 95, 407–424. [2673, 2675]

(2015), “Statistical Causality from a Decision-Theoretic Perspective,” Annual Review of Statistics and Its Application, 2, 273–303. [2673]

Didelez, V. (2018), “Defining Causal Mediation with a Longitudinal Mediator and a Survival Outcome,” Lifetime Data Analysis, 25, 593–610. [2682]

Ding, P., and Lu, J. (2017), “Principal Stratification Analysis Using Principal Scores,” Journal of the Royal Statistical Society, Series B, 79, 757–777. [2671, 2673]

Ding, P., Geng, Z., Yan, W., and Zhou, X.-H. (2011), “Identifiability and Estimation of Causal Effects by Principal Stratification with Outcomes Truncated by Death,” Journal of the American Statistical Association, 106, 1578–1591. [2681]
