Received: 12 August 2020 | Revised: 1 December 2020 | Accepted: 5 February 2021

DOI: 10.1002/pst.2104

WILEY

MAIN PAPER

Principal stratum strategy: Potential role in drug development

Björn Bornkamp[1] | **Kaspar Rufibach**[2] | **Jianchang Lin**[3] | **Yi Liu**[4] | **Devan V. Mehrotra**[5] | **Satrajit Roychoudhury**[6] | **Heinz Schmidli**[1] | **Yue Shentu**[7] | **Marcel Wolbers**[2]

[1]Clinical Development and Analytics, Novartis, Basel, Switzerland

[2]Methods, Collaboration, and Outreach Group (MCO), Department of Biostatistics, Hoffmann-La Roche Ltd, Basel, Switzerland

[3]Statistical & Quantitative Sciences (SQS), Takeda Pharmaceuticals, Cambridge, Massachusetts, USA

[4]Nektar Therapeutics, San Francisco, California, USA

[5]Clinical Biostatistics, Merck & Co., Inc., North Wales, Pennsylvania, USA

[6]Pfizer Inc., New York, New York, USA

[7]Merck & Co., Inc., Rahway, New Jersey, USA

Correspondence
Björn Bornkamp, Clinical Development and Analytics, Novartis, Basel, Switzerland.
Email: bjoern.bornkamp@novartis.com

Abstract

A randomized trial allows estimation of the causal effect of an intervention compared to a control in the overall population and in subpopulations defined by baseline characteristics. Often, however, clinical questions also arise regarding the treatment effect in subpopulations of patients, which would experience clinical or disease related events post-randomization. Events that occur after treatment initiation and potentially affect the interpretation or the existence of the measurements are called *intercurrent events* in the ICH E9(R1) guideline. If the intercurrent event is a consequence of treatment, randomization alone is no longer sufficient to meaningfully estimate the treatment effect. Analyses comparing the subgroups of patients without the intercurrent events for intervention and control will not estimate a causal effect. This is well known, but post-hoc analyses of this kind are commonly performed in drug development. An alternative approach is the principal stratum strategy, which classifies subjects according to their potential occurrence of an intercurrent event on both study arms. We illustrate with examples that questions formulated through principal strata occur naturally in drug development and argue that approaching these questions with the ICH E9(R1) estimand framework has the potential to lead to more transparent assumptions as well as more adequate analyses and conclusions. In addition, we provide an overview of assumptions required for estimation of effects in principal strata. Most of these assumptions are unverifiable and should hence be based on solid scientific understanding. Sensitivity analyses are needed to assess robustness of conclusions.

KEYWORDS

causal inference, estimand, intercurrent event, potential outcomes, randomization

1 | INTRODUCTION

One main concept of the E9(R1) guideline[1] by the International Council of Harmonization (ICH) is the notion of intercurrent events, defined as "... Events occurring after treatment initiation that affect either the interpretation or the

existence of the measurements associated with the clinical question of interest...." The ICH E9(R1) guideline outlines five strategies to acknowledge intercurrent events as part of the treatment effect/estimand of interest. The treatment policy strategy effectively makes the intercurrent event part of the treatment investigated. The composite and while-on-treatment strategies modify the variable/endpoint of interest to reflect the intercurrent event. The hypothetical strategy envisages a hypothetical scenario in which the intercurrent event does not occur. Finally, the principal stratum strategy, based on ideas introduced by Frangakis and Rubin,[2] defines a subpopulation of interest according to the potential occurrence of an intercurrent event on one or all treatments.

As part of a principal stratum strategy, the subpopulation of interest could for example be subjects who *would tolerate treatment if assigned to the test treatment*. In this case subpopulation membership on the test arm would be known. On the control arm however subpopulation membership is not observed and hence not known with certainty. Alternatively, the subpopulation of interest could be the patients who would tolerate both test and control treatment.

The principal stratum strategy has not been commonly used in clinical trials so far and is not uncontested, see Hernán and Scharfstein[3] and Scharfstein[4] or the discussion initiated earlier by Pearl[5]. First, it relates to a subpopulation of the overall trial population that is not identifiable with certainty (i.e., for some, or all, patients principal stratum membership is not observed and constitutes missing data). This may be perceived to render the obtained treatment effect estimate of limited interest from a direct practical perspective. Second, a principal stratum estimand relates to a question where one cannot rely on randomization anymore to ensure comparable baseline populations across treatment groups in the subpopulation of interest. Strong assumptions are typically needed to estimate this estimand. In the ICH E9(R1) guideline it is specifically mentioned that a run-in period may be an effective design feature to robustly identify a target population defined by a specific clinical event (and thus estimate a principal stratum effect). The use of these designs might however be limited to special situations.

For these reasons, one might be tempted to generally challenge the relevance of the principal stratum strategy in drug development.

In this paper we would like to illustrate with examples that many relevant scientific questions in drug development can be addressed with the principal stratum strategy. Often, these questions do not correspond to the primary endpoint in the specific trial, but they increase the scientific understanding of the treatment effect in relevant subpopulations, and may impact approval decisions and labeling.

The outline of this paper is as follows. In Section 2 we will provide a review of potential outcomes and principal stratum estimands. In Section 3 we review examples from drug development practice, where the question of interest can be framed to be of principal stratum type. Section 4 then reviews analysis methods and assumptions citing existing literature and outlines a R[6] implementation. The paper ends with a discussion in Section 5.

2 | INTRODUCTION TO POTENTIAL OUTCOMES AND PRINCIPAL STRATUM ESTIMANDS

The term *principal stratum* was first introduced by Frangakis and Rubin[2] (see also Mealli and Mattei[7] for a rather recent review on principal stratification) and originates from the causal inference literature under the potential outcome approach (see Hernán and Robins[8] and Imbens and Rubin[9] for introductions). In this section we will introduce potential outcomes, a central idea of causal inference, which are important to formulate principal stratum estimands. Note that the other estimand strategies in the ICH E9(R1) guideline can also be formulated using potential outcomes.[10] We illustrate potential outcomes with an example: Let Z be the binary indicator for treatment ($Z = 1$ corresponding to the test treatment and $Z = 0$ corresponding to control) and Y be the outcome of interest. Assume a treating physician is deciding on the treatment to prescribe. Ideally she would make that decision based on knowledge on what the outcome for the patient would be if given the control treatment, $Y(Z = 0)$, abbreviated as $Y(0)$, and what the outcome would be under test treatment, $Y(Z = 1) = Y(1)$. In reality of course, neither $Y(0)$ and $Y(1)$ is known when assigning a treatment, and even after observation, for a given patient, only one of the potential outcomes $Y(0)$ or $Y(1)$ can be observed. So, even after observation of Y one cannot be sure if the correct decision was made for this particular patient: Individual causal effects, that is, $Y(1) - Y(0)$, are not observed. On a population level, however, such "causal" statements can be made. One then targets the average causal effect $E(Y(1) - Y(0))$, where the expectation is taken with respect to the population of interest.

Statistical estimation of $E(Y(1) - Y(0))$ in a randomized trial can be performed based on the fact that treatment assignment is independent of any patient characteristic, so that $Y(1)$ and $Y(0)$ are independent of Z implying that

$$E(Y(1)-Y(0)) = E(Y(1)) - E(Y(0))$$
$$= E(Y(1)|Z=1) - E(Y(0)|Z=0)$$
$$= E(Y|Z=1) - E(Y|Z=0).$$

This means we can estimate the average causal effect by the difference in averages on the two arms, as the population of patients is comparable across the two treatment arms.

In an observational study the treatment decision between $Z = 0$ and $Z = 1$ might depend on further measured or unmeasured patient characteristics X, so that the patients who receive $Z = 1$ (for whom we observe $Y(1)$) might be systematically different from those patients who receive $Z = 0$ (for whom we observe $Y(0)$), so that $Y(1)$ and $Y(0)$ are not independent of Z. In this case $E(Y|\ Z = 1) - E(Y|\ Z = 0) \neq E(Y(1) - Y(0))$, because $E(Y(1)) \neq E(Y(1)|\ Z = 1)$ and $E(Y(0)) \neq E(Y(0)|\ Z = 0)$: The patients receiving $Z = 0$ are not representative of the overall population and similarly those receiving $Z = 1$ are not representative of the overall population. The value of potential outcomes from a notational perspective is that they allow to decouple the outcome Y from the actual treatment Z received.

Denoting by $Y(1)_i$ the potential outcome for a patient i and by $\mathcal{S}$ a population of patients, causal treatment effects are defined as a comparison of potential outcomes $\{Y(1)_i, i \in \mathcal{S}\}$ versus $\{Y(0)_i, i \in \mathcal{S}\}$ on a common set of units $\mathcal{S}$. A comparison of $\{Y(1)_i, i \in \mathcal{S}_1\}$ versus $\{Y(0)_i, i \in \mathcal{S}_2\}$ with $\mathcal{S}_1 \neq \mathcal{S}_2$ is not a causal effect.[11] A causal effect can thus be conceptualized as a comparison of outcomes "had everyone received treatment" versus outcomes "had everyone received control," see also Figure 1. This focus on causal effects is also present in the ICH E9(R1) guideline,[1] sect. A.3, where estimands are introduced as "... Central questions for drug development and licensing are to establish the existence, and to estimate the magnitude, of treatment effects: how the outcome of treatment compares to what would have happened to the same subjects under alternative treatment (i.e. had they not received the treatment, or had they received a different treatment)...."

To illustrate a main motivation of the principal stratum strategy we will consider a simple, generic example. Assume a randomized two-arm trial is planned, with an outcome Y assessed at week 12. Now assume that one is interested in the treatment effect in those patients that experience a specific post-randomization event of interest. Denote by $S = 1$ occurrence and by $S = 0$ absence of the post-randomization event. A naive analysis might subset the overall trial data to patients with $S = 1$ on both test and control arm and then perform the analysis of interest. The variable S is a post-randomization variable and an outcome influenced by treatment, that is, S depends on Z. This means for patients on the intervention arm we observe the potential outcome $S(Z = 1)$ and on control we observe the potential outcome $S(Z = 0)$. From this perspective the population of patients with $S(1) = 1$ and $S(0) = 1$ might be quite different. The naive analysis mentioned above is hence "breaking the randomization," as the patient populations on the compared arms can be different, and one is not comparing "like with like," and thus not estimating a causal effect. If we would numerically observe a treatment effect in such an analysis, we would not be sure whether the difference in outcome is due to the difference in treatment, or due to the difference in the compared populations.

The idea of principal stratum estimands is to stratify patients based on their potential outcomes $S(0)$, $S(1)$ for *all* treatments. In the case of a binary post-randomization event S and two treatments one can hence define four strata

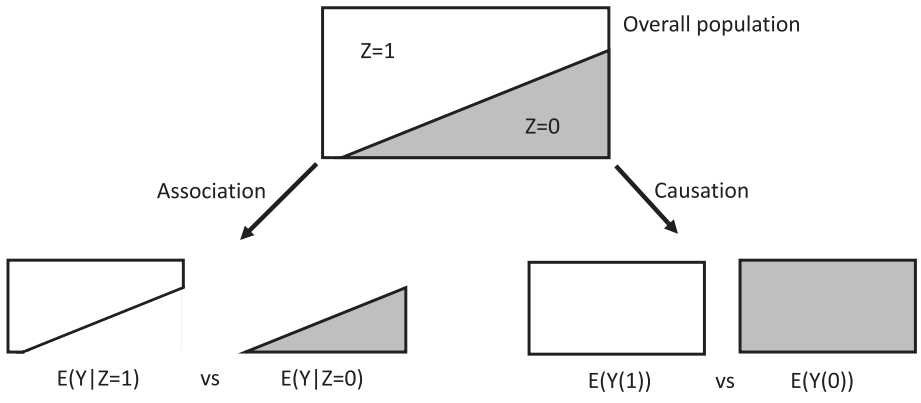

FIGURE 1 Difference between association and causation (adapted based on Hernán and Robins,[8] fig. 1.1)

	$S(0) = 1$	$S(0) = 0$
$S(1) = 1$	$\{S(1) = 1\} \cap \{S(0) = 1\}$	$\{S(1) = 1\} \cap \{S(0) = 0\}$
$S(1) = 0$	$\{S(1) = 0\} \cap \{S(0) = 1\}$	$\{S(1) = 0\} \cap \{S(0) = 0\}$

TABLE 1 Principal strata defined by the potential outcomes $S(0)$ and $S(1)$

based on both potential outcomes, see Table 1. Every patient falls into one particular of the four strata. Causal interpretations are made possible by the fact that membership to a principal stratum is not affected by treatment assignment.

In the described setting, patients that would experience the event under either treatment would have $S(1) = 1$ and $S(0) = 1$ (i.e., the top-left cell in Table 1). Alternatively of interest could be those patients that experience the event under treatment $S(1) = 1$, which is the union of the two cells $\{S(1) = 1\} \cap \{S(0) = 1\}$ and $\{S(1) = 1\} \cap \{S(0) = 0\}$ in Table 1.

Contrary to the naive analysis this stratification leads to a causal effect: We now stratify the population according to the same rule on treatment and control arm. The actual *identification* of the subpopulation corresponding to the stratum/strata of interest is generally not possible, not even after observing the outcome Y and the post-randomization event S in a given trial. For patients on the intervention arm we observe $S(1)$, but not $S(0)$, and vice versa for patients on the control arm.

Based on our experience (see also the examples in Section 3) one is often interested in the group of patients that experience the post-randomization event under one treatment (union of two principal strata, or a row in Table 1), for example, the stratum with $S(1) = 1$. Then in one arm patients in the stratum can be identified, but not on the other arm. Generally, assumptions are required for estimation of the treatment effect in the stratum/strata of interest. One could argue that the naive analysis (that subsets based on the observed S on both treatment arms) is estimating the treatment effect in the principal stratum $\{S(1) = 1\} \cap \{S(0) = 1\}$ under the assumption that $S(1) = S(0)$, that is, occurrence of the post-randomization event is not treatment related. Viewing this naive analysis within potential outcome notation reveals the implicit assumption underlying the analysis, which might often be quite strong and rarely justified.

While often one is primarily interested only in a subset of the overall trial population (one principal stratum or a union of strata), it is good practice to evaluate and report results also for the complementary group(s) or all strata (or union of strata) if the model allows to extract that information.

A complication arises when the outcome Y is not a measurement assessed at a specific timepoint, but of time-to-event type, such as death for overall survival (OS). In this situation the event itself can constitute a competing risk for occurrence of the intercurrent event (i.e., after observing the main event of interest, patients would no longer be at risk for experiencing the intercurrent event). In these situations particular care is needed to define the principal stratum of interest as well as the analysis strategy. Naive analyses conditioning on observed intercurrent event occurrence in this situation would then not only compare non-randomized populations but may also suffer from immortal bias, as patients for which we observe the intercurrent event are "immortal" until that timepoint.

In the causal inference literature the principal stratum approach has been controversially discussed in particular with respect to its relationship to mediation analysis, see for example Pearl.[5] In the latter, one tries to disentangle the overall effect via direct and indirect effects (mediated via the intercurrent event or not). In the language of the ICH E9(R1), mediation analysis can be interpreted as targeting a hypothetical estimand. See also the comparisons between principal stratum and mediation approaches in VanderWeele[12] on a conceptual level, and Bacchini et al.[13] on a practical example. In this paper, we will focus on principal stratum estimands, not the least because this concept has been proposed as one of five strategies to address intercurrent events in the ICH E9(R1) guideline.

3 | EXAMPLES

In this section we discuss scientific questions of interest in drug development that can be formulated as principal stratum estimands. For a discussion how to position these questions in the broader drug development landscape we refer to Section 5. In this section we will not discuss analysis strategies or assumptions that would allow estimation. We return to this aspect in Section 4. Table 2 gives an overview of the examples.

3.1 | Multiple sclerosis

Multiple sclerosis (MS) is an auto-immune disease of the central nervous system characterized by relapses, with varying symptoms, for example, visual deficits, cognitive and motor impairment. Multiple sclerosis typically starts out with a

TABLE 2 Examples of principal stratum estimands discussed in this section

Example	Scientific question	Primary endpoint	Intercurrent event	Stratum of interest
Multiple sclerosis	Treatment effect on confirmed disability progression in the subpopulation of relapse-free patients	Time to confirmed disability progression	Post-randomization relapse	Patients who would be relapse-free under both treatments
Treatment effect in early responders	Predict treatment effect on long-term primary endpoint based on early biomarker-type readout	Time-to-event	Biomarker value above or below a pre-specified threshold	Patients who would respond early under treatment vs. those that would not
Antidrug antibodies (ADA) for targeted oncology drugs	Do patients that develop ADAs on either arm still benefit from the drug?	Time-to-event	Development of antidrug antibodies because of receiving treatment	Patients who would be ADA+ under treatment
Impact of exposure on OS	Do patients with insufficient exposure have lower treatment effect?	Time-to-event	Exposure below a pre-specified threshold	Patients with low vs. non-low exposure under treatment
Prostate cancer prevention	Assess effect of treatment to prevent prostate cancer on severity of prostate cancer among those men who would be diagnosed with prostate cancer regardless of their treatment assignment	Time-to-event	Getting prostate cancer	Patients who get prostate cancer irrespective of treatment

phase where patients have relapses, but fully recover after the relapses (relapsing remitting form of MS, RRMS). Then the disease transitions to a phase where patients have a continuous disease progression where relapses are less common, and patients often do not fully recover from these, leading to increased disability (secondary progressive MS, SPMS). The typical primary endpoint in RCTs for SPMS is time to confirmed disability progression. In Magnusson et al.[14] EXPAND (NCT01665144, Kappos et al.[15]), a large placebo-controlled trial of siponimod in patients with SPMS, is discussed. The primary objective of the trial was to show efficacy of siponimod versus placebo in terms of time to confirmed disability progression. The endpoint was achieved, but the question was raised whether a treatment effect would also be present in patients that would not experience relapses. As siponimod is known to prevent relapses this is a non-trivial question to answer.

In this setting the intercurrent event S is post-randomization relapse, and Magnusson et al.[14] considered estimation of the treatment effect in patients that would not relapse under both siponimod and placebo, that is, in the stratum $\{S(0) = 0\} \cap \{S(1) = 0\}$. The probability of disability progression was assessed at a specific timepoint, so that the outcome Y is binary. The estimand of interest here was taken as the risk ratio

$$\mathrm{RR} := \frac{E(Y(1)|S(1) = 0, S(0) = 0)}{E(Y(0)|S(1) = 0, S(0) = 0)}.$$

3.2 | Treatment effect in early responders

Biomarkers or early readouts can be useful to investigate whether an investigational medicine works as intended on a biological level. In some situations, it is realistic to assume that patients, whose post-randomization short term biomarker levels indicate that they do not sufficiently respond to the drug, are also unlikely to respond on clinically relevant long term outcomes, such as time-to-event.

One recent example is in the cardiovascular area. Inflammation has been identified as playing a key role in atherosclerosis and cardiovascular disease. The CANTOS outcomes trial in prevention of cardiovascular events (NCT01327846, Ridker et al.[16]) investigated treatment with canakinumab, an anti-inflammatory agent, against placebo,

both on top of standard of care. The primary outcome was the time to major adverse cardiovascular event (MACE) and significant. In this specific case the biomarker of interest is a downstream inflammatory marker, high sensitivity c-reactive protein (hs-CRP), where lower values indicate less inflammation. Interest here was in determination of the treatment effect for patients that, 3 months after start of treatment with canakinumab, were able to lower hs-CRP below a specific target level. 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. Vice versa, patients that achieve the threshold have a larger treatment effect.

Another example is in oncology, where tumor size shrinkage is a measurement that can be assessed early. Again, for patients with a lack of tumor shrinkage it is less likely that those benefit from the treatment on longer term survival outcomes. In pharmacometrics so-called *tumor growth inhibition* (TGI) metrics have gained popularity. Such models are drug-independent and attempt to link tumor response and baseline prognostic factors to a time-to-event endpoint such as OS. In these models, tumor response is quantified by extracting a summary statistics from a longitudinal model of tumor size. The goal of these TGI analyses is to "predict" OS survival functions and induced effects of treatment based on such summary statistics (see, e.g., Han et al.[17]). Again, one could consider the treatment effect (in terms of OS) in patients that achieve a specific favorable tumor metric shortly after treatment start.

Determining the potential long-term treatment effect for a patient based on a short-term read-out, such as, for example, hs-CRP or TGI, can be useful information: Depending on the therapeutic setting and drug mechanism it might support the decision on treatment modifications after treatment start.

Let S denote the event of achieving an early readout value (i.e., hs-CRP or TGI) either (1) lower than a target level or (2) achieving a certain percent decrease with respect to the patient's baseline value, at a short time $\tilde{t}$ after start of treatment.

Interest focuses on comparing $Y(1)$ and $Y(0)$ in the stratum of patients with $S(1) = 1$, and contrasting this for example to the results for patients with $S(1) = 0$. Depending on the questions one could also be interested in the subpopulation of patients with $S(0) = 1$ and contrast results to those with $S(0) = 0$. Effect measures of interest can be based on the survival functions

$$U_1(t) := P(Y(1) > t | S(1) = 1) \text{ and } U_0(t) := P(Y(0) > t | S(1) = 1),$$

For example, event probabilities at a time $t^* > \tilde{t}$:

$$\delta(t^*) = U_1(t^*) - U_0(t^*)$$

or a time-averaged version $\int_0^{t^*} \delta(t)dt = E[\min(Y(1), t^*) - \min(Y(0), t^*)]$, the difference in restricted mean survival times.[18]

An important point to consider in these situations (as discussed in Anderson et al.[19]) is that response on the early read-out might simply act as a marker for prognostically favorable patients and thus not modify the treatment difference versus the control treatment itself. For example, comparing $Y(1)$ for patients with $S(1) = 1$ versus those with $S(1) = 0$, does not allow for a statement on the treatment effect (which is a contrast involving *Y(1)* and *Y(0)*).

Another challenge is that, depending on the time point $\tilde{t}$ of the measurement of the post-randomization marker, some events related to Y might already have happened. We discuss this general point later in Section 4.7.

3.3 | Antidrug antibodies for targeted oncology drugs

In oncology, an increasing number of targeted anticancer agents and immunotherapies are of biological origin.[20] These biological drugs may trigger immune responses that lead to the formation of antidrug antibodies (ADAs). 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.[21] ADA positivity (ADA+) is triggered by treatment, appears post-randomization and has the potential to affect the interpretation of the outcome. It can thus be considered an intercurrent event in the language of the ICH E9(R1) guideline. Note that in an RCT it can well be that a biologic drug is only administered in the test but not the control arm, that is, by construction ADAs can only form in the intervention arm. To make things concrete, assume that our outcome of interest Y is again a time-to-event endpoint, for example, OS. The intercurrent event S is occurrence of an ADA at a fixed milestone time point $\tilde{t}$ after randomization, for example, $\tilde{t} = 3$ weeks.

A relevant clinical question for the intercurrent event of ADA-positivity is whether ADA+ patients still benefit from the drug.

One way to answer the above clinical questions is to assess the effect of the randomized treatment in those patients that would be ADA+ under treatment, that is, we are interested in the treatment effect in the stratum $\{S(1) = 1\}$. This is the union of the two strata $\{S(1) = 1\} \cap \{S(0) = 1\}$ and $\{S(1) = 1\} \cap \{S(0) = 0\}$ in Table 1. The effect can then again be quantified via U_1 and U_0 introduced in Section 3.2.

Enrico et al.[22] give an overview of the issue of ADAs in the class of immune checkpoint blockers and re-analyze data of drug trials, by ADA status. They define ADA-positivity by "patient has ever been ADA+ during the observation period." However, as discussed in the previous section, naive analyses defining groups through a post-randomization event will lead to (i) the comparison of non-comparable population on the treatment groups and (ii) in this example also to immortal bias (see also, e.g., Anderson et al.,[19,23] Anderson,[24] Walraven et al.[25] and the discussion later in Section 4.7). Here, ADA+ patients were not at risk, and thus "immortal," of experiencing the outcome event between trial entry and the occurrence of ADA positivity. How much a causal conclusion based on the analysis in Enrico et al.[22] is justified is thus unclear. In the context of TGI metrics discussed in Section 3.2 this has been brought up in Mistry[26] as well.

3.4 | Impact of exposure on OS

The *Trastuzumab for Gastric Cancer* (ToGA) trial was a 1:1 Phase 3 RCT comparing chemotherapy versus chemotherapy + trastuzumab in patients with gastric or gastro-oesophageal junction cancer with over-expression of the HER2 protein.[27] Five hundred eighty-four patients entered the primary analysis. A post hoc exploratory analysis of OS by trastuzumab exposure in the intervention arm of ToGA was also performed,[28] where exposure was defined as trough minimum concentration, $C_{\min}$, at steady state in Cycle 1. Clearly, $C_{\min}$ is a post-randomization variable. The authors observed that patients with $C_{\min}$ values in the lowest quartile of the $C_{\min}$ distribution appeared to have shorter OS duration compared with other quartiles. In order to explore whether other (baseline) factors than exposure could contribute to the shorter observed OS in the lowest quartile group, further analyses of baseline patient characteristics by $C_{\min}$ quartile were performed. The conclusion was that "...it is unclear whether the lower OS is due to low drug concentration or to disease burden." In a follow-up analysis authors by the FDA in Yang et al.[29] evaluated the treatment effect in patients in the lowest $C_{\min}$ quartile (in the chemotherapy + trastuzumab arm) by appropriately matching these patients with patients in the chemotherapy only arm, to achieve covariate balance for key baseline covariates. Although not explicitly described in the potential outcomes framework, this approach implicitly targets a principal stratum estimand with S being an indicator for $C_{\min}$ below a given threshold after Cycle 1 on the test treatment, so that we are again in the situation of Section 3.2. This analysis together with further exposure-response analyses based on the ToGA data then triggered initiation of a fully-powered open-label RCT, HELOISE, that evaluated standard versus high dose trastuzumab,[30] for the identified subgroup of patients. This case-study illustrates the clinical importance of principal stratum estimands might have on a drug development program, even if not explicitly mentioned in the Yang et al.[29] paper.

3.5 | Prostate cancer prevention trial

The prostate cancer prevention trial (PCPT, Thompson et al.[31]), a double-blind RCT, randomized 18,882 men aged 55 years or older to finasteride or placebo. The trial convincingly showed that men randomized to finasteride had lower rates of prostate cancer. However in Thompson et al.[31] it was noted that among patients who developed prostate cancer after randomization, those randomized to finasteride had a statistically higher risk of high-grade prostate cancer compared to those randomized to placebo. The question of interest here is therefore assessing 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, see Lu et al.[32] for a very nice discussion. Severity was measured using the Gleason score, an ordered categorical variable that assigned integer values 2–10, with 10 being the most severe. To make things concrete, Z is the indicator of being randomized to finasteride, S is the indicator of getting prostate cancer, and Y is the Gleason score.

Interest thus focuses on the distribution functions of the two potential outcomes $Y(0)$ and $Y(1)$ in the stratum of those patients who get prostate cancer irrespective of treatment assignment, that is, $\{S(0) = 1\} \cap \{S(1) = 1\}$. Lu et al.[32]

describe how to estimate this effect and how to statistically test equality of distribution functions of the two potential outcomes.

In terms of results, naively looking at the distribution of Gleason scores in both arms suggests that those who got cancer in the finasteride arm had higher Gleason scores. This naive analysis however did not account for potential post-randomization selection bias due to differences among treatment arms in patient characteristics of cancer cases or differential biopsy grading associated with finasteride-induced reductions in prostate volume.[33] A subsequent sensitivity analysis based on principal stratification[34] accounting for these two potential sources of selection bias cast doubt on results from the aforementioned naive analysis. Indeed, a more recent report based on long-term follow-up of PCPT patients[35] has concluded that *The early concerns regarding an association between finasteride and an increased risk of high-grade prostate cancer have not been borne out*.

3.6 | Further examples

Targeting a principal stratum estimand has also been suggested for a variety of further examples and we sketch and reference some of these below.

A further application of principal stratum is discussed by Lou et al.[36,37] in the context of clinical equivalence studies, where intent-to-treat analyses are not considered "conservative" and per-protocol analyses are also commonly performed as primary analysis. Here the authors propose to use protocol adherence for S, and interest is in the stratum of patients that would adhere under both treatment and control. This corresponds to an estimand which is in spirit similar to the aims of the per-protocol analysis. In that direction the ICH E9(R1) guideline in Section A.5.3 also critically discusses per-protocol analyses, but also argues against handling of protocol deviations as one intercurrent event (protocol deviations may not necessarily be intercurrent events, and vice versa), and thus suggests a more granular handling of the different intercurrent events that may underlie protocol non-adherence.

Uemura et al.[38] propose to use principal stratum estimands for assessing quality of life in face of an intercurrent event that might happen before the assessment of quality of life. Typically in this case naive analyses are performed that ignore the intercurrent event.

In the context of schizophrenia, Larsen and Josiassen[39] are interested in the treatment effect on a continuous outcome Y in patients that would comply if treated with the test treatment, that is, the effect in the stratum $\{S(1) = 1\}$ in Table 1. They propose a new estimator for this setting.

Akacha et al.[40] suggest the tripartite estimand approach for characterizing the treatment effect in the overall population. They suggest reporting three numbers (i) non-adherence due to safety, (ii) non-adherence due to lack-of-efficacy and (iii) the effect in adherers. Estimand (iii) corresponds to a principal stratum strategy, see also Qu et al.[41] for a concrete application of this approach in a diabetes trial.

In the context of COVID-19 vaccine development a vaccine-induced dampening of disease severity might be clinically relevant (see U.S. Food and Drug Administration[42]), that is to reduce the *severity* of COVID-19 disease among the subset of those who will become infected despite vaccination.

4 | ANALYSIS METHODS AND ASSUMPTIONS

Estimates for principal stratum estimands rely on the validity of assumptions beyond randomization. In the literature, a variety of possible assumptions have been suggested and the choice of the most appropriate particular set of assumptions will depend on the context of each specific case. The literature on analysis methods for principal stratum estimands is vast and an extensive review is beyond the scope of this article. In what follows we provide a selected overview of commonly utilized assumptions. In addition, we discuss possible sensitivity analyses in Section 4.6 and considerations specific to principal stratum estimands for time-to-event endpoints in Section 4.7.

To allow readers to implement some of the analyses presented below, we have developed a R[6] markdown,[43,44] see "Data Availability Statement" for the link. The file generates an exemplary clinical trial data-set containing potential outcomes Y and S as well as a categorical covariate X mimicking the case study in Section 3.3, and provides explicit code for some of the analyses described below. Finally, a sensitivity analysis is sketched.

4.1 | SUTVA and consistency

Most approaches described rely on the stable unit treatment values assumption (SUTVA), which entails that (i) the potential outcomes for any patient do not change with the treatment assigned to other patients (no interference) and (ii) there are no multiple versions of treatment. An example where (i) is violated is in the area of infectious diseases: Depending on the context, whether or not an individual may get infected will depend on whether other individuals are vaccinated. Part (ii) implies that treatment needs to be well-defined so that potential outcomes corresponding to a defined treatment are equal to what is observed in the trial. This is sometimes also called "consistency" assumption.[45] In addition, most approaches utilize the fact that there is an ignorable treatment assignment mechanism, as is common in pharmaceutical RCTs.

4.2 | Identification bounds

One stream of literature tries to avoid utilizing assumptions. This means that typically no concrete estimate can be provided but only identification bounds for the parameters of interest, see, for example, Zhang and Rubin[46] and Chiba and VanderWeele.[47] The estimation problem then focuses on estimation of these identification boundaries, for which also confidence intervals can be provided. Often these bounds might be quite wide and might not provide useful information, but this depends on the specific data situation. Refinements of boundaries using, for example, covariate information were discussed in Grilli and Mealli,[48] Long and Hudgens,[49] Mealli and Pacini.[50]

4.3 | Monotonicity and exclusion-restriction

Two possible "nonparametric" assumptions to utilize are the monotonicity assumption and the exclusion-restriction assumption.[51] The monotonicity assumption states that $S(0) \geq S(1)$ (or alternatively $S(1) \geq S(0)$ depending on the situation). This means for a patient with $S(0) = 0$ observed we would know that $S(1) = 0$, so that the bottom-left stratum in Table 1 would be empty. This assumption allows to estimate the principal stratum probabilities. The monotonicity assumption may in some situations be scientifically very plausible, but is not verifiable based on observed data. It however implies that $P(S(0) = 1) \geq P(S(1) = 1)$, an assumption that can be assessed in a RCT.

The exclusion restriction assumption states that for patients in the strata $\{S(0) = 0\} \cap \{S(1) = 0\}$ and $\{S(0) = 1\} \cap \{S(1) = 1\}$ one assumes $Y(0) = Y(1)$, that is, there would be no treatment effect in the strata of those experiencing (or not experiencing) the post-randomization event under either treatment. Formulated alternatively, randomization has no impact for those subjects for whom treatment has no effect on S.[52]

Note that both assumptions make statements on the relationship of potential outcomes across treatment and control. As potential outcomes across treatment and control are never observed jointly, these type of assumptions are typically not verifiable and can be called "across-worlds" assumptions. The naming "across-worlds" comes from the fact that this assumption could only be verified if two "parallel" worlds could be observed jointly, in which a patient would receive control in one world and treatment in the other parallel world.

In the context of the multiple sclerosis example of Section 3.1 these assumptions together would allow identification of the estimand of interest. But while a monotonicity assumption can well be justified based on earlier data, the exclusion-restriction assumption is not plausible, as the estimand of interest is the treatment effect in the stratum with no relapses $\{S(0) = 0\} \cap \{S(1) = 0\}$.

4.4 | Joint models

In Frangakis and Rubin[2] a generic likelihood is described for estimation of a principal stratum effect. This entails a model for the potential outcomes given the principal stratum membership: $Y(0), Y(1) \mid S(1), S(0)$ and additionally the principal stratum membership $S(0), S(1)$ itself is modeled. Multiplying the likelihoods of both models together, implies a joint model for Y and S. Unobserved potential outcomes are then treated as missing data in Frangakis and Rubin[2] and integrated out to define the likelihood. While covariates are not specifically mentioned, including them in the model for the principal stratum membership or the outcome is straightforward. As noted in Frangakis and Rubin[2] a

unique maximum likelihood estimate generally does not exist (even asymptotically for "infinite" sample size). Further assumptions are needed, which typically involve statements on the joint distribution of the potential outcomes across treatment and control (across-world assumptions).

In a Bayesian setting, prior assumptions on the model parameters can be quantified in terms of prior distributions and in this case, as long as proper priors are used, also posterior inference is possible. This idea goes back at least to Imbens and Rubin[53] and was implemented for estimation in the example in Section 3.1. There a "soft" version of the monotonicity assumption was used, by specifying an informative prior distribution for the corresponding principal stratum proportion to be close to 0. This allows for sensitivity analyses through varying the informativeness of the prior distribution. While the model by Magnusson et al.[14] in Section 3.1 did not include covariates to model the principal stratum membership, this is possible (see, e.g., Michela et al.,[13] Zhang et al.,[54] Hirano et al.,[55] Frumento et al.,[56] Mattei et al.,[57] Mealli et al.[58] in a Bayesian setting).

It is often plausible to assume that covariates might influence principal stratum membership, so that inference will get more precise. This approach can also be coupled with additional assumptions like monotonicity and exclusion restriction, which will improve identification of the underlying inference problem. For example, Jo and Stuart[59] and Stuart and Jo[60] employ the exclusion restriction assumption in addition to a parametric assumption in the context of linear regression, while using the EM-algorithm for ML estimation (extension to a time-to-event outcome using instrumental variable approaches is discussed in MacKenzie et al.[61] and Martinussen et al.[62]). A general challenge, independent of whether a Bayesian or frequentist inference paradigm is used, is that inference will depend on the parametric assumptions of the underlying joint models, as well as the covariates used.

In general, when making a parametric assumption on the distribution of the data, the parameters are identified by the identification of parametric mixture models, and an EM-type algorithm can be used for statistical inference. Note however that the likelihood function may display pathological behavior, and standard frequentist inference tools, like bootstrap cannot be used. For details, further discussion and references see Ding and Li[63] sect. 2.3.2.

4.5 | Principal ignorability

The last type of assumption we discuss is *conditional independence*. In the context of principal strata this is often called *principal ignorability* (PI), see Ding and Lu[64] and Feller et al.[65] for some recent references. Here, separate models are specified for Y and S, resulting in approaches that are very similar to propensity score approaches in observational data analyses. For example, in the early responder example in Section 3.2 the estimand of interest was

$$P(Y(1) > t | S(1) = 1) - P(Y(0) > t | S(1) = 1),$$

where Y was an event time and S early response. Contrary to $P(Y(1) > t \mid S(1) = 1)$, estimation of $P(Y(0) > t \mid S(1) = 1)$ is not straightforward, because $Y(0)$ and $S(1)$ are not jointly observed in the same patient in a RCT. For patients on treatment that are biomarker responders, that is, $S = 1$, the control outcome $Y(0)$ is unobserved, while for patients on the control arm the biomarker response status on treatment $S(1)$ is not observed.

Now, PI states that conditional on baseline covariates (i.e., confounders) X the $Y(0)$ and $S(1)$ are independent: $Y(0) \perp S(1) \mid X$, so the covariates X should include those that explain both $Y(0)$ and $S(1)$ to the extent that they can be considered independent. This means once the covariates X are known, $S(1)$ provides no further information on $Y(0)$ and vice versa, that is, the distributional equality

$$p(Y(0) | X, S(1)) = p(Y(0) | X)$$

holds. The benefit from this assumption is that it allows modeling of $Y(0)$ (or $S(1)$) just based on X, the unobserved outcome does not need to be included in the model. Based on this, weighting approaches based on propensity scores can, for example, be used as follows. First, model the probability that $S(1) = 1$ on the treatment arm depending on X, for example, using logistic regression. Then, use the predicted probabilities as weights for patients on the control arm (see Stuart and Jo[60] and Bornkamp and Bermann[66] among many others). The same model could also be used in a multiple imputation approach, that is, imputing $S(1)$ for patients on the control arm. An even simpler approach may often be standardization.[8] Depending on the outcome distribution also plain regression adjustment for X in the outcome model can be utilized to estimate a principal stratum effect under the PI assumption. Finally, matching is also feasible.

The propensity score literature extensively discusses the pros and cons of the different analysis techniques, see, for example, Austin.[67]

The main assumption of PI is that X contains all variables that potentially confound $Y(0)$ and $S(1)$ (no unmeasured confounding). As $Y(0)$ and $S(1)$ cannot be jointly observed, this assumption is hence across-worlds and not verifiable. Its plausibility needs to be considered on a case-by-case basis.

One practically relevant and important question is how to decide on the covariates X to use. Here an important point is to adjust for all confounders that make the potential outcomes of post-randomization event occurrence and the final outcome independent. So, formally only covariates that confound the two outcomes should be adjusted for, that is, formally one should not include covariates that help predicting the intercurrent event but have no impact on the outcome. The discussion on which variables to utilize is very similar to the discussion in observational data settings, where one tries to find predictors of both treatment and outcome. A helpful recent overview is provided by Persson et al.,[68] who study the nonparametric setting.

While the approach by Larsen and Josiassen[39] also utilizes covariates, it utilizes an assumption quite different to principal ignorability: It appears that utilized covariates are required to have no effect on the outcome, beyond their effect mediated via compliance.[69]

4.6 | Sensitivity analyses

Because essentially all analysis strategies targeting principal stratum estimands require strong assumptions, sensitivity analyses should be performed, depending on how strong the scientific rationale for the utilized assumptions is. The proposed sensitivity analyses are often specific to the specific assumptions utilized so that consequently a number of different sensitivity analyses approaches exist, see, for example, Ding and Lu,[64] Shepherd et al.,[70] Schwartz et al.,[71] Ding and VanderWeele[72] and references cited therein. In a Bayesian approach sensitivity analyses are often relatively straightforward, as for example discussed in Magnusson et al.[14]

In the setting of clinical equivalence trials, Lou et al.[36,37] were interested in the treatment effect in the patient that adhere under both treatments. They utilize an idea of Chiba and VanderWeele Tyler[47] to express the estimand of interest in terms of the naive per-protocol effect with a bias term. Estimation of the target estimand is then done by the naive per protocol effect and varying the bias term within reasonable bounds in a tipping point analysis. In addition they propose to test for equivalence of the proportions of protocol adherence under both treatments as a co-primary endpoint in clinical equivalence studies.

Applying methodology initially proposed in Lu et al.[32] and Gilbert et al.[73] in the PCPT example from Section 3.5 apply a sensitivity analysis to assess robustness of the naive analysis that does not account for potential post-randomization selection bias. They make the distribution function of the potential outcome in the placebo patients depend on a parameter β that can be interpreted as follows: given someone got prostate cancer in the placebo arm, for a one-unit increase in Gleason score, the odds that they would have gotten prostate cancer had they been randomized to finasteride arm multiplicatively increases by $\exp(\beta)$. Plotting the estimated relative effect between the two potential outcomes (or the p-value as in Lu et al.[32] if interest focuses on hypothesis testing) against this parameter β then allows an assessment of the dependency of the conclusion on the amount of selection bias.

For assumptions related to principal ignorability tipping point analyses can be used to assess the sensitivity of assumptions to the underlying conclusions. Here $Y(0)$ or $S(1)$ would be used as potential predictor of $S(1)$ or $Y(0)$ on top of X. As the effect of $Y(0)$ on $S(1)$ or the effect of $S(1)$ on $Y(0)$ cannot be estimated based on data, these effects would need to be varied in a sensitivity analysis. Another important sensitivity analysis related to principal ignorability is to vary the set of confounders X utilized.

4.7 | Special considerations for time-to-event endpoints

As discussed earlier, event-driven endpoints require special considerations, as the primary event in some situations might be a competing risk to observing the intercurrent event status, potentially leading to immortal bias when utilizing naive analyses. If the primary event is death this is obvious, but this situation might also occur, for example, when the primary event triggers a stop of the treatment and the intercurrent event can only happen while on treatment. Then intercurrent event cannot be observed after stop of treatment.

Assume we are interested in the subpopulation with $S(1) = 1$ and the outcome Y is of time-to-event type, further let $T_S(1)$ be the potential event time for occurrence of $S(1)$. Then in the situation described above the event $S(1) = 1$ implies $T_S(1) < Y(1)$, so that implicitly the stratum of interest would be $\{T_S(1) < Y(1)\}$.

When the intercurrent event status $S(1)$ is observed for every unit at a fixed time $\tilde{t}$, the occurrence of the primary event Y before $\tilde{t}$ would make observation of $S(1)$ impossible so that observation of $S(1)$ implies $Y(1) > \tilde{t}$. The stratum $S(1) = 1$ would then implicitly be defined as $\{S(1) = 1\} \cap \{Y(1) > \tilde{t}\}$.

In both situations the stratum is no longer only described by $S(1)$, but also by $Y(1)$. As the main comparison of interest is between $Y(0)$ and $Y(1)$ and the principal stratum itself is now also defined in terms of the observed outcome $Y(1)$, it becomes more challenging to find realistic assumptions that would allow estimation of principal stratum effects.

When interest focuses on $\{S(1) = 1\} \cap \{Y(1) > \tilde{t}\}$ it is easier to find plausible assumptions when $\tilde{t}$ is "small," that is, very few events Y are expected before time $\tilde{t}$. This could, for example, be fulfilled in the early responder, exposure, or ADA example in Sections 3.2, 3.3, 3.4, when the intercurrent event status can be identified early. Depending on the specific situation (e.g., if Y measures non-fatal events), in the three examples above it might also be possible to assess $S(1)$ even after the event $Y(1)$ has already happened, so that one would not necessarily be in the situation discussed in this section, where the event and intercurrent events are competing risks.

This problem is also discussed in some detail in Mattei et al.[74] in the context of treatment switching, where the event (death in their considered case) is a competing risk to treatment switch.

5 | DISCUSSION

We believe that there are a number of relevant questions in drug development that can be formulated as principal stratum estimands. These are often not related to the primary objective of the trial, but can still play an important role to characterize how the drug works in relevant subpopulations defined by different post-randomization events.

That these type of questions occur in regulatory interactions and are considered worthwhile to provide guidance upon, can seen from the anticancer guidance issued by the European Medicines Agency,[75] which has a dedicated section (7.6.5) on "Analyses based on a grouping of patients on an outcome of treatment." The MS example discussed in Section 3.1 is also available as Public Assessment Report of the European Medicines Agency.[76] In addition, of course, the ICH E9(R1) estimand working group considered the principal stratum strategy as important enough to list it as one of the five intercurrent event strategies.

Questions about the impact of clinical events such as exposure, response, or safety events like ADA on the outcome of interest have always been relevant in drug development. However, although often criticized in the literature, simple analyses such as comparing subgroups based on a post-randomization event are not uncommon in an attempt to answer such questions. Causal effects were, at least implicitly, claimed from such analysis. Even though the formal idea of principal stratum estimands had been proposed in the causal inference literature two decades ago, the explicit uptake of these methods in the drug development community has so far been low. While there are examples of analyses that appropriately would target principal stratum estimands (e.g., Yang et al.[29]), explicit use of principal stratum estimands has been limited. Exceptions exist, for example, to assess efficacy on a post-infection endpoint in a vaccine trial in Mehrotra et al.[77] (where a principal stratum estimand has been used for a primary endpoint). We believe and hope this will change with the principal stratum approach being prominently mentioned as a strategy in the ICH E9(R1) guideline. The advantage of adopting the principal stratum strategy for these questions is that it provides a clear inferential target. Having an inferential target is crucial to assess the adequacy of assumptions or specific analyses.

Even more generally, in our experience approaching traditional analyses with a potential outcome mindset often allows to make implicit assumptions of traditional analyses more transparent, as in the example discussed in Section 2.

The type of assumptions typically required for identification of principal stratum estimands are quite strong and usually unverifiable. While similar type of unverifiable assumptions have long been used in drug development, for example, missing-at-random or independent censoring assumptions, the impact might be stronger here as assumptions are not only used to "impute" missing responses for a potentially small subset of the overall trial population, but depending on the data situation and the type of assumption might drive inference. However, availability of a clear inferential target though based on unverifiable assumptions has to be traded off against "naive" analyses whose causal interpretation is unclear, if not to say invalid.

Design considerations can help to make the utilized assumptions more plausible, for example, to explicitly consider at design stage, which baseline and post-baseline covariates may be associated with stratum membership, and

subsequent collection of these measurements in the trial. While not feasible in all situations, also cross-over designs (similar to the run-in period discussed in ICH E9(R1)) may facilitate estimation of principal stratum estimands under potentially weaker assumption.

In general we think however that (i) utilized assumptions need to be motivated by clinical or scientific insights and (ii) that sensitivity analyses need to be performed for any analysis targeting a principal stratum estimand. While sensitivity analyses for certain assumptions, for example, monotonicity, have been proposed in the literature and we tried to review some ideas for further sensitivity analyses in this paper, we believe there is a need for further developments and practical guidance in this area.

ACKNOWLEDGMENTS

This paper has been written within the industry working group *estimands in oncology*, which is both, a *European special interest group* "Estimands in oncology," *sponsored by PSI and European Federation of Statisticians in the Pharmaceutical Industry (EFSPI)* and a scientific working group of the biopharmaceutical section of the American Statistical Association. Details are available on www.oncoestimand.org. We are grateful for feedback of working group colleagues, as well as from Kelly Van Lancker and Fabrizia Mealli on earlier versions of this manuscript. We would also like to thank two anonymous reviewers for helpful comments.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The markdown file discussed in Section 4 is available as a github repository: https://github.com/oncoestimand/princ_strat_drug_dev.git. The direct link to the markdown file is: https://oncoestimand.github.io/princ_strat_drug_dev/princ_strat_example.html.

ORCID

Björn Bornkamp https://orcid.org/0000-0002-6294-8185
Kaspar Rufibach https://orcid.org/0000-0002-2634-1167
Devan V. Mehrotra https://orcid.org/0000-0002-0316-7362
Satrajit Roychoudhury https://orcid.org/0000-0003-4001-3036

REFERENCES

1. International Council of Harmonization. Addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials E9(R1); 2019. https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf
2. Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics*. 2002;58:21-29.
3. Hernán MA, Scharfstein DO. Cautions as regulators move to end exclusive reliance on intention to treat. *Ann Internal Med*. 2018;168:515-516.
4. Scharfstein DO. A constructive critique of the draft ICH E9 Addendum. *Clin Trial*. 2019;16:375-380.
5. Pearl J. Principal stratification—a goal or a tool? *Int J Biostat*. 2011;7:20.
6. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria; 2020.
7. Mealli F, Mattei A. A refreshing account of principal stratification. *Int J Biostat*. 2012;8:1-19.
8. Hernán MA, Robins JM. *Causal Inference: What If*. Boca Raton: CRC Press; 2020.
9. Imbens GW, Rubin DB. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge: Cambridge University Press; 2015.
10. Lipkovich I, Ratitch B, Mallinckrodt CH. Causal inference and estimands in clinical trials. *Stat Biopharm Res*. 2020;1:54-67.
11. Rubin DB. Causal inference using potential outcomes: Design, modeling, decisions. *J Am Stat Assoc*. 2005;100:322-331.
12. VanderWeele TJ. Simple relations between principal stratification and direct and indirect effects. *Stat Prob Lett*. 2008;78:2957-2962.
13. Baccini M, Mattei A, Mealli F. Bayesian inference for causal mechanisms with application to a randomized study for postoperative pain control. *Biostatistics*. 2017;18:605-617.
14. 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;38:4761-4771.
15. Kappos L, Bar-Or A, Cree Bruce AC, et al. Siponimod versus placebo in secondary progressive multiple sclerosis (EXPAND): a double-blind, randomised, phase 3 study. *Lancet*. 2018;391:1263-1273.
16. Ridker PM, Everett BM, Thuren T, et al. Antiinflammatory therapy with canakinumab for atherosclerotic disease. *New Engl J Med*. 2017;377:1119-1131.

17. Han K, Claret L, Piao Y, et al. Simulations to predict clinical trial outcome of bevacizumab plus chemotherapy vs. chemotherapy alone in patients with first-line gastric cancer and elevated plasma VEGF-A. *CPT Pharmacomet Syst Pharmacol*. 2016;5:352-358.
18. Royston P, Parmar Mahesh KB. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Med Res Method*. 2013;13:152.
19. Anderson JR, Cain KC, Gelber RD. Analysis of survival by tumor response and other comparisons of time-to-event by outcome variables. *J Clin Oncol*. 2008;26:3913-3915.
20. Brummelen EMJ, Ros W, Wolbink G, Beijnen JH, Schellens JHM. Antidrug antibody formation in oncology: clinical relevance and challenges. *Oncologist*. 2016;21:1260-1268.
21. Moussa Ehab M, Panchal Jainik P, Moorthy Balakrishnan S, et al. Immunogenicity of therapeutic protein aggregates. *J Pharmaceut Sci*. 2016;105:417-430.
22. Enrico D, Paci A, Chaput N, Karamouza E, Besse B. Anti-drug antibodies against immune checkpoint blockers: impairment of drug efficacy or indication of immune activation? *Clin Cancer Res*. 2019;26:787-792.
23. Anderson JR, Cain KC, Gelber RD. Analysis of survival by tumor response. *J Clin Oncol*. 1983;1:710-719.
24. Anderson JR. Commonly misused approaches in the analysis of cancer clinical trials. In: John C, ed. *Handbook of Statistics in Clinical Oncology*. 1st ed. New York: Dekker; 2001:525-542.
25. Walraven C, Davis D, Forster AJ, Wells GA. Time-dependent bias was common in survival analyses published in leading clinical journals. *J Clin Epidemiol.* 2004;57:672-682.
26. Mistry HB. Time-dependent bias of tumor growth rate and time to tumor regrowth. *CPT Pharmacomet Syst Pharmacol*. 2016;5:587.
27. Bang Y-J, 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;376:687-697.
28. 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;73:737-747.
29. 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;53:160-166.
30. Shah MA, Xu RH, 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;35:2558-2567.
31. Thompson IM, Goodman PJ, Tangen CM, et al. The influence of finasteride on the development of prostate cancer *N. Engl J Med*. 2003; 349:215-224.
32. Lu X, Mehrotra DV, Shepherd BE. Rank-based principal stratum sensitivity analyses. *Stat Med.* 2013;32:4526-4539.
33. Lucia M, Scott EJI, Goodman Phyllis J, et al. Finasteride and high-grade prostate cancer in the prostate cancer prevention trial. *J Nat Cancer Inst*. 2007;99:1375-1383.
34. Shepherd BE, Redman MW, Ankerst DP. Does finasteride affect the severity of prostate cancer? A causal sensitivity analysis. *J Am Stat Assoc*. 2008;103:1392-1404.
35. Goodman PJ, Tangen CM, Darke AK, et al. Long-term effects of finasteride on prostate cancer mortality. *New Engl J Med*. 2019;380: 393-394.
36. Lou Y, Jones MP, Sun W. Assessing the ratio of means as a causal estimand in clinical endpoint bioequivalence studies in the presence of intercurrent events. *Stat Med*. 2019;38:5214-5235.
37. Lou Y, Jones MP, Sun W. Estimation of causal effects in clinical endpoint bioequivalence studies in the presence of intercurrent events: noncompliance and missing data. *J Biopharmaceut Stat*. 2019;29:151-173.
38. Uemura Y, Taguri M, Kawahara T, Chiba Y. Simple methods for the estimation and sensitivity analysis of principal strata effects using marginal structural models: application to a bone fracture prevention trial. *Biomet J*. 2019;61:1448-1461.
39. Larsen KG, Josiassen MK. A new principal stratum estimand investigating the treatment effect in patients who would comply, if treated with a specific treatment. *Stat Biopharmaceut Res*. 2020;1:29-38.
40. Akacha M, Bretz F, Ruberg S. Estimands in clinical trials–broadening the perspective. *Stat Med*. 2017;36:5-19.
41. Qu Y, Fu H, Luo J, Ruberg SJ. A general framework for treatment effect estimators considering patient adherence. *Stat Biopharmaceut Res*. 2020;12:1-18.
42. U.S. Food and Drug Administration. *Guidance for Industry*: *Development and Licensure of Vaccines to Prevent COVID-19*; 2020.
43. Xie Y, Allaire JJ, Garrett G. *R Markdown: The Definitive Guide*. Boca Raton, FL: Chapman and Hall/CRC; 2018.
44. Allaire JJ, Xie Y, Jonathan MP, et al. *rmarkdown*: *Dynamic Documents for R*. R package version 2.3; 2020.
45. VanderWeele TJ, Hernán MA. Causal inference under multiple versions of treatment. *J Causal Infer*. 2013;1:1-20.
46. Zhang JL, Rubin DB. Estimation of causal effects via principal stratification when some outcomes are truncated by death. *J Educ Behav Stat*. 2003;28:353-368.
47. Chiba Y, VanderWeele TJ. A simple method for principal strata effects when the outcome has been truncated due to death. *Am J Epidemiol*. 2011;173(7):745-751.
48. Grilli L, Mealli F. Nonparametric bounds on the causal effect of university studies on job opportunities using principal stratification. *J Educ Behav Stat*. 2008;33:111-130.
49. Long DM, Hudgens MG. Sharpening bounds on principal effects with covariates. *Biometrics*. 2013;69:812-819.

50. Mealli F, Pacini B. Using secondary outcomes to sharpen inference in randomized experiments with noncompliance. *J Am Stat Assoc*. 2013;108:1120-1131.
51. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *J Am Stat Assoc*. 1996;91:444-455.
52. Joffe MM, Small D, Hsu C-Y, et al. Defining and estimating intervention effects for groups that will develop an auxiliary outcome. *Stat Sci*. 2007;22:74-97.
53. Imbens GW, Rubin DB. Bayesian inference for causal effects in randomized experiments with noncompliance. *Ann Stat*. 1997;25(1): 305-327.
54. Zhang JL, Rubin DB, Mealli F. Likelihood-based analysis of causal effects of job-training programs using principal stratification. *J Am Stat Assoc*. 2009;104:166-176.
55. Hirano K, Imbens GW, Rubin DB, Zhou X-H. Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*. 2000;1:69-88.
56. Frumento P, Mealli F, Pacini B, Rubin DB. Evaluating the effect of training on wages in the presence of noncompliance, non-employment, and missing outcome data. *J Am Stat Assoc*. 2012;107:450-466.
57. Mattei A, Li F, Mealli F. Exploiting multiple outcomes in Bayesian inference for causal effects with intermediate variables. *Ann Appl Stat*. 2013;7:2336-2360.
58. Mealli F, Pacini B, Stanghellini E. Identification of principal causal effects using additional outcomes in concentration graphs. *J Educ Behav Stat*. 2016;41:463-480.
59. Jo B, Stuart EA. On the use of propensity scores in principal causal effect estimation. *Stat Med*. 2009;28:2857-2875.
60. Stuart EA, Jo B. Assessing the sensitivity of methods for estimating principal causal effects. *Stat Method Med Res*. 2015;24:657-674.
61. MacKenzie TA, Løberg M, O'Malley AJ. Patient centered hazard ratio estimation using principal stratification weights: application to the norccap randomized trial of colorectal cancer screening. *Observ Stud*. 2016;2:29.
62. Martinussen T, Vansteelandt S, Tchetgen Tchetgen Eric J, Zucker DM. Instrumental variables estimation of exposure effects on a time-to-event endpoint using structural cumulative survival models. *Biometrics*. 2017;73:1140-1149.
63. Ding P, Li F. Causal inference: A missing data perspective. *Stat Sci*. 2018;33:214-237.
64. Ding P, Lu J. Principal stratification analysis using principal scores. *J Roy Stat Soc B*. 2017;79:757-777.
65. Feller A, Mealli F, Miratrix L. Principal score methods: Assumptions, extensions, and practical considerations. *J Educ Behav Stat*. 2017; 42:726-758.
66. Bornkamp B, Bermann G. Estimating the treatment effect in a subgroup defined by an early post-baseline biomarker measurement in randomized clinical trials with time-to-event endpoint. *Stat Biopharmaceut Res*. 2020;12:19-28.
67. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res*. 2011;46:399-424.
68. Persson E, Häggström J, Waernbaum I, De Luna X. Data-driven algorithms for dimension reduction in causal inference. *Comput Stat Data Anal*. 2017;105:280-292.
69. Shepherd BE, Gilbert PB, Dupont CT. Sensitivity analyses comparing time-to-event outcomes only existing in a subset selected post-trandomization and relaxing monotonicity. *Biometrics*. 2011;67:1100-1110.
70. Schwartz S, Fan L, Reiter JP. Sensitivity analysis for unmeasured confounding in principal stratification settings with binary variables. *Stat Med*. 2012;31:949-962.
71. Ding P, VanderWeele TJ. Sensitivity analysis without assumptions. *Epidemiology*. 2016;27:368-377.
72. Chiba Y, VanderWeele TJ. A simple method for principal strata effects when the outcome has been truncated due to death. *Am J Epidemiol*. 2011;173:745-751.
73. Gilbert PB, Bosch RJ, Hudgens MG. Sensitivity analysis for the assessment of causal vaccine effects on viral load in HIV vaccine trials. *Biometrics*. 2003;59:531-541.
74. Mattei A, Mealli F, Ding P. Assessing Causal Effects in the Presence of Treatment Switching Through Principal Stratification; 2020. https://export.arxiv.org/abs/2002.11989
75. European Medicines Agency. Guideline on the evaluation of anticancer medicinal products in man; 2017. https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-evaluation-anticancer-medicinal-products-man-revision-5_en.pdf
76. European Medicines Agency. Committee for Medicinal Products for Human Use. Mayzent: Assessment Report; 2019.
77. Mehrotra DV, Xiaoming L, Gilbert PB. A comparison of eight methods for the dual-endpoint evaluation of efficacy in a proof-of-concept HIV vaccine trial. *Biometrics*. 2006;62:893-900.

How to cite this article: Bornkamp B, Rufibach K, Lin J, et al. Principal stratum strategy: Potential role in drug development. *Pharmaceutical Statistics*. 2021;20:737–751. https://doi.org/10.1002/pst.2104