Received: 28 February 2022 | Revised: 3 June 2022 | Accepted: 27 July 2022

DOI: 10.1002/pst.2260

WILEY

MAIN PAPER

Sensitivity analyses for the principal ignorability assumption using multiple imputation

Craig Wang[1] | Yufen Zhang[2] | Fabrizia Mealli[3,4] | Björn Bornkamp[1]

[1]Department of Analytics, Novartis Pharma AG, Basel, Switzerland

[2]Department of Analytics, Novartis Pharmaceuticals Corp, East Hanover, New Jersey, USA

[3]Department of Statistics, Computer Science and Applications, Florence Center for Data Science, University of Florence, Florence, Italy

[4]Economics Department, European University Institute, Florence, Italy

Correspondence
Craig Wang, Department of Analytics, Novartis Pharma AG, Basel, Switzerland.
Email: craig.wang@novartis.com

Abstract

In the context of clinical trials, there is interest in the treatment effect for subpopulations of patients defined by intercurrent events, namely disease-related events occurring after treatment initiation that affect either the interpretation or the existence of endpoints. With the principal stratum strategy, the ICH E9(R1) guideline introduces a formal framework in drug development for defining treatment effects in such subpopulations. Statistical estimation of the treatment effect can be performed based on the principal ignorability assumption using multiple imputation approaches. Principal ignorability is a conditional independence assumption that cannot be directly verified; therefore, it is crucial to evaluate the robustness of results to deviations from this assumption. As a sensitivity analysis, we propose a joint model that multiply imputes the principal stratum membership and the outcome variable while allowing different levels of violation of the principal ignorability assumption. We illustrate with a simulation study that the joint imputation model-based approaches are superior to naive subpopulation analyses. Motivated by an oncology clinical trial, we implement the sensitivity analysis on a time-to-event outcome to assess the treatment effect in the subpopulation of patients who discontinued due to adverse events using a synthetic dataset. Finally, we explore the potential usage and provide interpretation of such analyses in clinical settings, as well as possible extension of such models in more general cases.

KEYWORDS

causal inference, estimand, principal stratum, subgroup analysis, survival analysis

1 | INTRODUCTION

The ICH E9(R1) addendum proposes a framework for the formulation of scientific questions and treatment effects a clinical trial should address.[1] A central notion introduced in the addendum is that of intercurrent events, which are "events occurring after treatment initiation that affect either the interpretation or the existence of the measurements associated with the clinical question of interest". Examples of such events include treatment discontinuation, initiating concomitant rescue medication and treatment switching. The addendum proposes five strategies to address intercurrent events at the estimand level.[2] In this paper, we focus on the principal stratum strategy. This strategy, motivated by Frangakis and Rubin,[3] deals with intercurrent events by focusing on the subpopulation(s), where an intercurrent event would (or would not) occur in one or all treatment arms. Bornkamp et al.[4] recently reviewed different examples in drug

development where the strategy may be of interest. In some situations, certain design features like run-in phases and re-randomization approaches may be utilized to allow for estimation of principal stratum effects: one would observe the occurrence of an intercurrent event in a first phase (post-baseline), and then re-randomize patients in a second phase to assess the treatment effect in a subpopulation defined by the intercurrent event. If such designs are not feasible, then some assumptions that typically cannot be directly verified are required for identification of the principal strata and estimation of the treatment effect within, such as the exclusion restriction assumption or assumptions aimed at reducing the total number of principal strata (e.g., monotonicity; see References 5–7 for some criticism of the principal stratum strategy[8]; provide insights on the origin and interpretation of principal stratum estimands and related analysis strategies). In the presence of unverifiable assumptions, as it is typical in any observational setting, it is important to conduct sensitivity analyses to evaluate the robustness of conclusions to deviations from such assumptions. For example, Lou et al.[9] recently used the principal stratum strategy and combined it with a sensitivity analysis in bioequivalence trials. The authors used results by Chiba and Vanderweele[10] to identify the principal stratum effect and then vary a sensitivity analysis parameter within reasonable bounds in the spirit of a tipping point analysis. Similar bounds (and sensitivity analyses) for principal stratum effects are reviewed in Richardson et al.[11] However, these sensitivity analyses are not developed for situations where covariates are utilized to identify the principal stratum membership.

A particular assumption that has been put forward for the estimation of principal stratum estimands is principal ignorability.[12–14] Under principal ignorability, one assumes that all baseline covariates (confounders) are observed conditional on which membership in (some of) the principal strata and (some of) the potential outcomes of interest are independent. Different versions of principal ignorability exist that are sufficient for the identification of different effects. With the principal ignorability assumption, the causal effect can be estimated after identifying subpopulations with the same distribution of covariates across treatment arms. A common approach for balancing covariates across treatment arms is through propensity score analysis using either matching, weighting or regression adjustment. In analogy to instrumental variables, Jiang and Ding[15] additionally used auxiliary variables in the estimation of causal effects. These auxiliary variables may take the form of secondary outcomes as in Mealli and Pacini[16] or covariates for which specific conditional independence assumptions hold.

In this paper, we model stratum memberships using baseline covariates with multiple imputation approaches (see also Westreich et al. and Ding and Li[17,18] for the connection of the missing data and causal inference literature). This ensures that patients in the same stratum across treatment arms will have similar distributions of the baseline covariates.

The principal ignorability assumption is similar to the ignorability or no unmeasured confounders assumption in observational data analyses. For the latter situation, a number of sensitivity analyses have been proposed starting with Cornfield et al.[19] For example, Rosenbaum and Rubin[20] estimated the treatment effect assuming existence of an unobserved latent binary confounder. Their sensitivity analysis consists of varying the effect that the unobserved confounder may have on the outcome and on treatment assignment (after adjusting for confounders), as well as varying the distribution of the confounder. A more recent approach in the same spirit is proposed by Ding and VanderWeele,[21] which does not assume confounders of a particular form. Stuart and Jo,[22] in the context of estimating the complier average causal effect, proposed models based on the exclusion restriction and the principal ignorability assumptions and used one versus the other as sensitivity analysis. In the context of the principal ignorability assumption, Ding and Lu[13] proposed a principal scoring weighted approach to estimate causal effect and a number of sensitivity analyses but did not consider them for time-to-event outcomes or the implied conditional independence assumption directly. Furthermore, no guidance was provided on the choice of sensitivity parameter values and its meaningful interpretations, both of which are important in practice. Additional discussion on sensitivity analysis with discrete principal strata is provided in Mattei et al.[23]

In this paper, we focus on time-to-event endpoints, which are commonly used in oncology trials. Our sensitivity analysis is motivated by the structural form of the conditional independence assumption implied by principal ignorability. We consider a situation where the intercurrent event can only occur for patients randomized to the investigational treatment arm. Since the occurrence of the primary event used for assessing the treatment effect is often a competing risk to that of the intercurrent event, the interpretation of analysis results could be challenging and warrants extra caution.

In the following, Section 2 reviews a case study, which is motivated from an oncology clinical trial. Section 3 then introduces our proposed strategies for sensitivity analyses based on the principal ignorability assumption. Section 4 evaluates the operating characteristics of such approaches using simulations and in Section 5 we illustrate our approach

implemented on a synthetic dataset that was motivated by the same case study. Section 6 concludes the paper with summary and discussions.

2 | MOTIVATING EXAMPLE

We consider a phase III randomized, double blind, placebo-controlled oncology trial. Patients were randomized using a 1:1 ratio to receive either investigational treatment consisting of a new treatment (TRT) plus existing standard of care (SOC) or control treatment consisting of placebo (PBO) plus SOC. Each patient starts with both treatments drug components and if applicable, they were allowed to discontinue both components, or only TRT or PBO while still remaining on SOC. The primary endpoint in this study was progression-free survival (PFS), which is defined as the time from randomization until disease progression per RECIST v1.1 criteria[24] or death due to any cause, whichever occurs earlier. The occurrence of disease progression triggers a discontinuation of the study treatment, and patients are then transitioned to the next line of anti-cancer therapy at the discretion of treating physicians.

Figure 1 illustrates some patient journey examples in this study. Patient 1 has disease progression before end of study without any intercurrent events. Patient 2 discontinues TRT or PBO prematurely (e.g., due to adverse events) and continues on SOC alone until end of study. Patient 3 has no progression or death, and no intercurrent events before the analysis cut-off time. Patient 4 discontinues TRT or PBO prematurely and continues on SOC alone for a while, then discontinues SOC and starts new anti-cancer therapy, then shows disease progression. Patient 5 discontinues TRT or PBO prematurely and continues on SOC alone and is then lost to follow-up. In this trial, patients are more likely to discontinue one component of the study treatment (TRT or PBO) and to continue on SOC alone (like patient 2 or 5 in Figure 1). There were only a few patients like patient 4 who discontinued both TRT/PBO and SOC and started a new anti-cancer therapy before disease progression.

In oncology trials, the intercurrent event of premature treatment discontinuation (either one or both components of the study treatment) is commonly handled using the treatment policy strategy in the primary estimand of PFS (see Reference 25). In order to accommodate this in the estimand definition, the treatment attribute was defined to be the intended study treatment plus any next line anti-cancer therapies as needed.

In the motivating example, the treatment effect in terms of PFS results for the primary estimand claimed both statistical and clinical significance for the overall study population. However the proportion of patients who discontinued TRT due to adverse event (AE) in the treatment arm was notably higher (25%) than the proportion of patients who discontinued PBO due to AE in the control arm (<5%). Thus, there was interest to better understand whether patients who discontinued TRT due to AE still benefited from the new treatment (i.e., comparing to the situation, where they would have only received SOC).

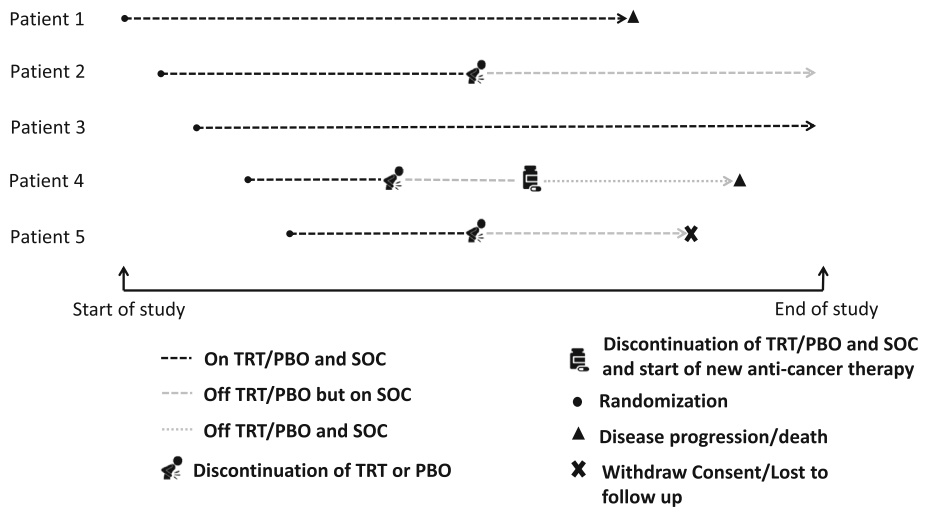

FIGURE 1 Illustration of patient journeys in example oncology trial

3 | METHODOLOGY

3.1 | Notation

Without loss of generality, we use PFS as an example of a time-to-event endpoint for the remainder of this paper. We use potential outcomes to define causal effects ([26,27]) and make the stable unit treatment value assumption, that is, the potential outcome of one patient is unaffected by the assigned treatment of another patient and there is no hidden variation of the treatments. Let us denote with $Y_i(z)$ and $C_i(z)$ the potential PFS event time and censoring time for the i^{th} patient under treatment assignment z, with $z = 0$ for the control arm and $z = 1$ for the treatment arm. Censoring here refers to patients that are followed up until the analysis cut-off time or until withdraw consent/lost to follow-up without experiencing the main event (disease progression or death). Further let $T_i(z)$ be the potential time to treatment discontinuation due to AE under treatment assignment z.

Specifically, our interest lies in the treatment effect (i.e., a comparison of $Y(1)$ vs. $Y(0)$) in the subpopulation of patients (principal stratum) who experience a treatment discontinuation due to AE when assigned to the treatment arm, that is, patients for whom

$$T_i(1) \leq \min\left(C_i(1), Y_i(1)\right). \tag{1}$$

Since it is common for oncology trials to continue collect progression data after premature treatment discontinuation (except for administrative reasons such as lost to follow-up or withdraw consent), it is reasonable to assume non-informative censoring, that is, $T(1), Y(1) \perp C(1)$. Let $S_i(1)$ be a binary variable that is equal to 1 if $T_i(1) \leq \min\left(C_i(1), Y_i(1)\right)$ and 0 otherwise. Hence, the principal stratum we are interested in is the subpopulation of patients with $S_i(1) = 1$.

From this definition of the principal stratum of interest, two aspects become obvious: (1) $S_i(1)$ cannot be observed on any of the patients assigned the control arm; it is therefore unclear which patients on the control arm would have fulfilled Equation (1) had they been randomized to the treatment arm; (2) to be part of the stratum requires $Y_i(1) \geq T_i(1)$, the event time $Y_i(1)$ is hence reasonably large for patients in this stratum. For patients who do not discontinue before experiencing either disease progression or death ($S_i(1) = 0$), $T_i(1)$ becomes undefined, as treatment discontinuation due to reasons other than disease progression cannot be observed after patient has already discontinued treatment due to disease progression. The stratum in Equation (1) will hence consist of patients that have diagnosis moderate enough to not immediately experience the PFS event, but are at the same time fragile from suffering a treatment discontinuation due to adverse event (before disease progression).

3.2 | Basic approach

To estimate the causal treatment effect on Y for patients in the principal stratum $S_i(1) = 1$, we need to estimate a contrast between the conditional distributions

$$p(Y(1)|S(1) = 1) \quad \text{and} \quad p(Y(0)|S(1) = 1). \tag{2}$$

While $p(Y(1)|S(1) = 1)$ can be estimated on the treatment arm, the distribution $p(Y(0)|S(1) = 1)$ can be estimated based on the principal ignorability assumption, see for example[28] for a detailed derivation in a similar situation.

A sufficient assumption for identifying the principal stratum effect of interest is a weak version of principal ignorability that requires that all baseline covariates X are measured such that $S(1) \perp Y(0)|X$. This implies that once we know X for a patient, $S(1)$ does not provide further information on the PFS event time under control arm (i.e., $Y(0)$), and vice versa. This assumption cannot be tested, as we cannot observe simultaneously $S(1)$ and $Y(0)$ for the same patient.

However if the assumption holds, we can easily construct models to identify the subset of patients on the treatment and the control arm such that the comparison of $Y(1)$ and $Y(0)$ can be made for patients with $S(1) = 1$. This can be done in one of two ways:

1. Identify the subset of patients on the control arm with $S(1) = 1$, that is, control arm patients that would have discontinued due to AE if they were randomized to the treatment arm

(1) Start with constructing a binary outcome model for $S(1)$ given X on the treatment arm.
(2) Use the model to impute $S(1)$ for patients on the control arm.
(3) Select all patients on the control arm with imputed $S(1) = 1$ as the comparator group for patients on the treatment arm with $S(1) = 1$ and evaluate the treatment effect.

2. Evaluate $Y(0)$ for patients on the treatment arm with $S(1) = 1$, that is, what their PFS would have been if they were randomized to the control arm
(1) Start with constructing a survival model for $Y(0)$ given X on the control arm.
(2) Use the model to impute $Y(0)$ for patients on the treatment arm with $S(1) = 1$.
(3) Take the imputed $Y(0)$ and observed $Y(1)$ to evaluate the treatment effect in patients on the treatment arm with $S(1) = 1$.

For each of these two approaches, multiple imputations could be utilized, that is, using covariates to multiply impute the missing $S(1)$ for patients on the control arm, or to multiply impute the missing $Y(0)$ for patients on the treatment arm. At the end, the overall treatment effect estimate can be obtained by combining the estimates from multiple imputations using Rubin's rule.[29] The discussion here is not restricted to any specific model or effect measure. However, for consistency it may be desirable to use the same statistical model and effect measure as the pre-specified primary analysis of the clinical trial.

The use of the principal ignorability assumption can be challenged in this application for several reasons. First it may not be feasible in practice to know and to capture all covariates X that make $S(1)$ and $Y(0)$ independent; this is a criticism that always applies to principal ignorability in general. In addition, an issue in the specific situation here is that patients need to have $T(1) \leq \min(C_i(1), Y_i(1))$ to be part of the principal stratum (i.e., have $S(1) = 1$). The term $\min(C_i(1), Y_i(1))$ is usually dominated by $Y(1)$, so that the principal ignorability assumption essentially implies conditional independence between $Y(0)$ and (a function of) $Y(1)$ given covariates. The paper[30] discusses that it can be unrealistic that $Y(1)$ does not provide any information on $Y(0)$ given covariates, which also renders the principal ignorability assumption in our setting questionable.

To overcome this challenge, we propose a joint model for sensitivity analysis that no longer relies on the principal ignorability assumption. Through the joint model, we impose conditional dependence between $S(1)$ and $Y(0)$ (and implicitly thus to some extent a conditional dependence between $Y(1)$ and $Y(0)$) such that analyses can be performed under different level of conditional dependency between $S(1)$ and $Y(0)$ by varying a sensitivity parameter. In a practical situation, such an analysis can be used to assess, whether, or how much, the main conclusions of an analysis are sensitive to the assumption of principal ignorability.

3.3 | Joint imputation model for sensitivity analysis

For purpose of imputation under a deviation of principal ignorability, we construct a joint model $Y(0), S(1)|X$, in which the conditional independence assumption under principal ignorability does not hold,

$$f(Y(0), S(1)|X) = f(Y(0)|S(1), X) f(S(1)|X). \tag{3}$$

We construct this joint model by defining separate models for $S(1)|X$ and then $Y(0)|S(1), X$. Here the model for $Y(0)|S(1), X$ depends on a sensitivity parameter that quantifies the association between $Y(0)$ and $S(1)$ given X. We propose to fit this model with Bayesian methods, which allows for straightforward imputation of the unobserved potential outcomes.

Different models with various flexibility could be utilized for $S(1)|X$ and $Y(0)|S(1), X$. For the ease of presentation we focus on parametric methods, specifically modeling $S(1)|X$ using a logistic regression with linear predictor $x_i'\alpha$. Note that only patients on the treatment arm provide information for $S(1)$ while the potential outcome $S(1)$ is missing for patients on the control arm. Similarly there are different approaches to model $Y(0)|S(1), X$. Here we use a Weibull model with density $p(y|\nu, \lambda) = \nu y^{\nu-1} \exp(-\lambda y^{\nu})$ and associated hazard rate $h(y) = \lambda \nu y^{\nu-1}$, where the scale parameter λ_i for patient i is modeled depending on covariates and $S_i(1)$ as

$$\lambda_i = \exp(x'_i\beta + \gamma S_i(1)). \tag{4}$$

Using this approach, only patients on the control arm provide information on $Y(0)$. The term γ is a sensitivity parameter that introduces dependency between $Y(0)$ and $S(1)$ and will be fixed during model fitting. When $\gamma = 0$, principal ignorability between $S(1)$ and $Y(0)$ given X holds. When $\gamma \neq 0$, the model for $Y(0)$ (fitted on data on the control arm) depends on $S(1)$, which is unobserved on the control arm and needs to be imputed. This can be resolved by fitting both models $S(1)|X$ and $Y(0)|S(1), X$ together in a joint Bayesian model setting.

While for the purpose of description we focused on a simple parametric Weibull model, the Appendix includes an example on how finite mixture of Weibull distributions can be used. These have recently been shown to provide adequate flexibility for time to event data in the context of oncology studies.[31]

Using Gibbs-sampling MCMC methods, the missing potential outcomes ($S(1)$ for patients on the control arm and $Y(0)$ for patients on the treatment arm) are automatically sampled from the corresponding full conditional distributions, given all observed and imputed data used in the model as well as the set of sampled model parameters. The posterior samples for $S(1)$ for all control arm patients can hence be used as part of a multiple imputation strategy.

The parameter γ plays a crucial role in this analysis. Selecting $\gamma < 0$ will create positive association between $S(1)$ and $Y(0)$ given X: Patients with $S(1) = 1$ on treatment have a lower hazard for PFS event on control, hence will get a larger imputed $Y(0)$ and vice versa for $\gamma > 0$. Information in this joint model also flows in the reverse direction: Patients with larger observed $Y(0)$ on the control arm are more likely to get $S(1) = 1$ imputed when $\gamma < 0$. Selecting $\gamma < 0$ will thus reduce the estimated treatment effect in the stratum of patients with $S(1) = 1$ (as patients with observed large $Y(0)$ on control are more likely to get $S(1) = 1$ imputed). Similarly for $\gamma > 0$ the estimated treatment effect will increase for patients with $S(1) = 1$.

3.3.1 | Choice of γ

A range of different values for γ should be investigated through the sensitivity analysis. As argued in Section 3.2, the value $\gamma = 0$ (where principal ignorability holds) may often not be the most plausible value for a given situation.

A range of plausible values can be proposed based on a reasonable assumption that $S(1)$ and $Y(1)$ should be more strongly associated than $S(1)$ and $Y(0)$ even after adjustment for X: $S(1)$ and $Y(1)$ happen "in the same world" with $S(1)$ having more direct association with $Y(1)$ as they can be observed concurrently for all patients in the treatment arm; while $S(1)$ and $Y(0)$ happen "in different worlds", that is either S(1) or Y(0) is observable for a specific patient, but not both of them. In other words, $S(1)$ is a potential outcome for patients in the placebo arm. One way to determine the association between $Y(1)$ and $S(1)$ given X, is to fit a Weibull model analogous to the one mentioned above, but now only to patients on the treatment arm, where $S(1)|X$ and $Y(1)|S(1), X$ are modeled with $\lambda_i = \exp(x'_i\beta + \eta S_i(1))$. The limits of a posterior credible interval for η could then be used as the range for γ. The range could be extended if the value $\gamma = 0$ (principal ignorability holds) was not covered by this interval.

An alternative fully Bayesian approach for sensitivity analysis could be to utilize a prior distribution for the parameter γ and fit this as part of the joint model. The caveat is that the joint model will only poorly inform γ, so that the prior for γ would need to be carefully chosen. A further alternative is Monte Carlo-based sensitivity analysis (see McCandless and Gustafson[32]), where γ is fixed during model fitting, but the inferences for given γ are averaged over sampled values from a pre-defined probability distribution for γ. However, we believe that it is insightful to present the results for a range of γ values rather than averaging the results.

3.3.2 | Choice of X

Based on the principal ignorability assumption, all confounders between $Y(0)$ and $S(1)$ need to be chosen. In practice, one should hence utilize the covariates affecting both $Y(0)$ and $S(1)$. For many diseases, covariates affecting $Y(0)$ (prognostic factors) are usually better established than covariates affecting $S(1)$. A general strategy is hence to include prognostic factors in the model for $Y(0)$ as long as they could potentially impact $S(1)$. Another more empirical approach is to include all covariates and possibly interactions that may affect either $Y(0)$ and $S(1)$ and then utilize variable selection (shrinkage) prior distributions, such as the horseshoe prior[33] to determine the final model.

3.3.3 | Model implementation

Markov Chain Monte Carlo (MCMC) sampling as implemented in JAGS[34] can be used to sample from the joint posterior of all model parameters and the unobserved potential outcomes. The full conditional distributions of the model parameters as well as the unobserved potential outcomes ($S(1)$ values for patients on control and $Y(0)$ values for patients on the treatment arm) are automatically sampled as part of the JAGS MCMC algorithm. Note the outcome $Y(1)$ is not used for fitting the imputation model. For the model parameters α, β, and ν, weakly informative prior distributions are used. An example Bayesian model and an alternative frequentist implementation are described in the Appendix.

4 | SIMULATIONS

We conduct a simulation study to evaluate the operating characteristics of the proposed joint model for different values of γ that cover scenarios when principal ignorability holds or does not hold. In the simulation study, we will also include a naive subpopulation analysis as a benchmark for the comparison.

4.1 | Simulation setup

We simulate clinical trials where patients are randomized using a 1:1 ratio to either treatment or control arm. The proposed joint imputation model is implemented with specific sensitivity parameters and compared with the naive subpopulation analysis, to evaluate the treatment effect for the strata of patients with treatment discontinuation due to AE.

Below are detailed steps to generate the data for each simulated trial. For each patient with treatment assignment Z we generate

1. Baseline covariates

$$X_1, X_2, U \sim N(0,1)$$

where X_1 and X_2 are observable and U is unobserved.

2. The binary status of treatment discontinuation due to AE on the treatment arm $S(1)$ from a binomial distribution with probability depending on X_1 and U as

$$S(1) \mid X_1, U \sim \text{Bin}\left(\text{logit}^{-1}(\alpha_0 + \alpha_1 X_1 + \alpha_2 U)\right)$$

3. The PFS event times under each arm $Y(1)$ and $Y(0)$ from exponential distributions based on X_1, X_2, and $S(1)$,

$$\begin{aligned} Y(1)|X_1, X_2, S(1) &\sim \text{Exp}(\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \delta + \eta I(S(1)=1))) \\ Y(0)|X_1, X_2, S(1) &\sim \text{Exp}(\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \gamma I(S(1)=1))) \end{aligned} \tag{5}$$

Whilst we have simulated Z, X_1, X_2, U, $S(1)$, $Y(0)$ and $Y(1)$ for each patient, for model fitting we only consider the 'observable' data, that is assuming $S(1)$ is only available for patients randomized to the treatment arm and each patient only has either $Y(0)$ or $Y(1)$ observed, depending on the treatment arm: $Y = Y(1)I(Z=1) + Y(0)I(Z=0)$.

The variable X_1 plays the role of a confounder, that is, it influences both $S(1)$ and $Y(0)$ and will lead to biased results if it is not adjusted in the joint model. The variable X_2 is not a confounder since it only impacts the outcome Y. The unobserved variable U is also not a confounder but it can help with predicting $S(1)$, that is, $\alpha_2 \neq 0$ will lead to worse predictions for $S(1)$. The parameter δ can be interpreted as the treatment effect for patients with $S(1) = 0$ and $\delta + \eta - \gamma$ as the treatment effect for patients with $S(1) = 1$ on the log hazard scale. This data-generating process implies principal ignorability $Y(0) \perp S(1)|X_1$ holds as long as $\gamma = 0$. If $\gamma \neq 0$, the value of $S(1)$ has a direct impact on $Y(0)$ such that $p(Y(0)|S(1), X_1) \neq p(Y(0)|X_1)$ and thus principal ignorability does not hold.

We consider 16 different scenarios that differ in terms of the number of patients in each trial ($N = 200$ or 400), the underlying principal stratum size (25% or 50% of total trial population), conditional dependency between $Y(0)$ and $S(1)$ indicating deviation from principal ignorability (γ being 0 or 1) and finally deviations due to unmeasured variable that predicts $S(1)$ (α_2 being 0 or 1). The remaining parameters are fixed as $\alpha_1 = 2$, $\beta_0 = -5$, $\beta_1 = 0.5$, $\beta_2 = -0.5$, and $\delta = \log(0.5)$ A larger value of the confounder X_1 thus implies a larger probability to have $S(1) = 1$ and also an increase in the hazard for the PFS event. The value of δ implies a treatment effect in terms of (conditional) hazard ratio being 0.5 for the patients with $S(1) = 0$. In addition, α_0 is set to achieve the specified underlying stratum size based on fixed α_1 and α_2. Similarly, given γ and δ, η is computed to achieve a fixed hazard ratio of 0.8 for the principal stratum of interest (patients with $S(1) = 1$), that is, $\eta = \log(0.8) + \gamma - \delta$. Table 1 shows different parameter combinations used in the simulation study.

We use Cox proportional hazard model (with and without covariate adjustment for X_1 and X_2) for the analysis of treatment effect.

For each scenario in Table 1, 1000 simulations of trials with sample size either 200 or 400 are performed with log hazard ratios evaluated for the principal stratum of $S(1) = 1$ using four different models:

1. Naive (model 1–2): Naive subpopulation analysis based on subset of patients from the treatment arm with $S(1) = 1$ and all patients from the control arm, and fitting a Cox model without any covariate adjustment to estimate the treatment effect. As a second approach (Naive + Cov) a Cox model is fitted that additionally adjusts for X_1 and X_2.
2. PI (assume Principal Ignorability holds; model 3): Joint imputation model taking patients from treatment arm with $S(1) = 1$ and patients from control arm with imputed $S(1) = 1$ using sensitivity parameter $\gamma = 0$. The prior distributions for the model parameters in the joint imputation model were selected weakly informative: $\nu \sim \Gamma(0.01, 0.01)$, $\alpha_i \sim N(0, \sigma^2 = 1000)$ and $\beta_i \sim N(0, \sigma^2 = 1000)$ for $i \in \{0, 1, 2\}$. For each imputation, a Cox model adjusting for X_1 and X_2 is used to estimate a log-hazard ratio. Results for the log-hazard ratios are at the end combined via the Rubin's rule to calculate a point estimate and its corresponding 95% confidence interval. In fact, the joint imputation model is not necessary in this setting, as no information flows across the logistic regression and Weibull model when $\gamma = 0$. Therefore, the imputations here could also be done by the logistic regression model alone.
3. NPI-EC (assume No Principal Ignorability, as an Extreme Case; model 4): The same joint imputation model is used as for PI, but now using a sensitivity parameter γ set to the point estimate $\widehat{\eta}$ estimated from $Y(1)$ and $S(1)$ using a Weibull model. We consider this to be an "extreme" case as the actual level of dependency between $Y(0)$ and $S(1)$ is likely to be smaller than that between $Y(1)$ and $S(1)$.

TABLE 1 Summary of parameter values used in the simulation study, with each scenario being repeated for $N = 200$ and 400

Scenario	δ	η	γ	Stratum size	α_0	α_1	α_2	β_0	β_1	β_2
1	−0.69	0.47	0.00	25%	−1.78	2.00	0.00	−5.00	0.50	−0.50
2	−0.69	0.47	0.00	25%	−1.91	2.00	1.00	−5.00	0.50	−0.50
3	−0.69	0.47	0.00	50%	−0.00	2.00	0.00	−5.00	0.50	−0.50
4	−0.69	0.47	0.00	50%	−0.00	2.00	1.00	−5.00	0.50	−0.50
5	−0.69	1.47	1.00	25%	−1.78	2.00	0.00	−5.00	0.50	−0.50
6	−0.69	1.47	1.00	25%	−1.91	2.00	1.00	−5.00	0.50	−0.50
7	−0.69	1.47	1.00	50%	−0.00	2.00	0.00	−5.00	0.50	−0.50
8	−0.69	1.47	1.00	50%	−0.00	2.00	1.00	−5.00	0.50	−0.50

When fitting the joint imputation models, JAGS version 4.3.0 is used with 10,000 MCMC samples, 5000 of them being run-in samples and thinning factor set to 5. The bias, coverage of 95% CI and mean squared errors (MSE) were obtained by comparing log(HR) estimates with its true value. Simulations were performed using a cluster of Linux nodes via R version 3.6.1.

4.2 | Simulation results

The estimated log hazard ratios are shown in Figure 2 for each scenario, and the corresponding coverage of 95% CI and MSE are shown in Figure 3 and Figure 4.

Based on the simulated scenarios, patients with $S(1) = 1$ are expected to have larger hazard rates (i.e., shorter PFS event times) when $\eta, \gamma > 0$. The naive analyses select patients with $S(1) = 1$ on the treatment arm and all of the patients on the control arm (with a mix of $S(1) = 0$ and $S(1) = 1$), hence it will underestimate the treatment effect (i.e., overestimate the hazard ratio) since the model does not take the 'correct comparator group of patients' from the control arm.

When the principal ignorability assumption holds ($\gamma = 0$), $Y(0)$ and $S(1)$ are conditionally independent such that all patients from the control arm can be taken as the comparator group in the covariates-adjusted model regardless of $S(1) = 0$ or $S(1) = 1$. Therefore, naive approach that adjusts for X_1 and X_2 and the PI method perform similarly well in terms of bias, coverage and MSE. On the other hand, the NPI-EC model consistently underestimates the hazard ratio (overestimates the treatment effect) in the principal stratum since it assumes the association between $Y(0)$ and $S(1)$ is the same as the estimated association between $Y(1)$ and $S(1)$. This assumption deviated from the data generation process (i.e., no association between $Y(0)$ and $S(1)$ conditional on X_1 when $\gamma = 0$). The estimated value $\widehat{\eta}$ (which in the data generation is greater than 0) is used for γ (which in truth is equal to 0). In such situations, the NPI-EC model is more likely to impute $S(1) = 1$ for control patients with worse outcome (smaller $Y(0)$) hence driving the hazard ratio estimates lower.

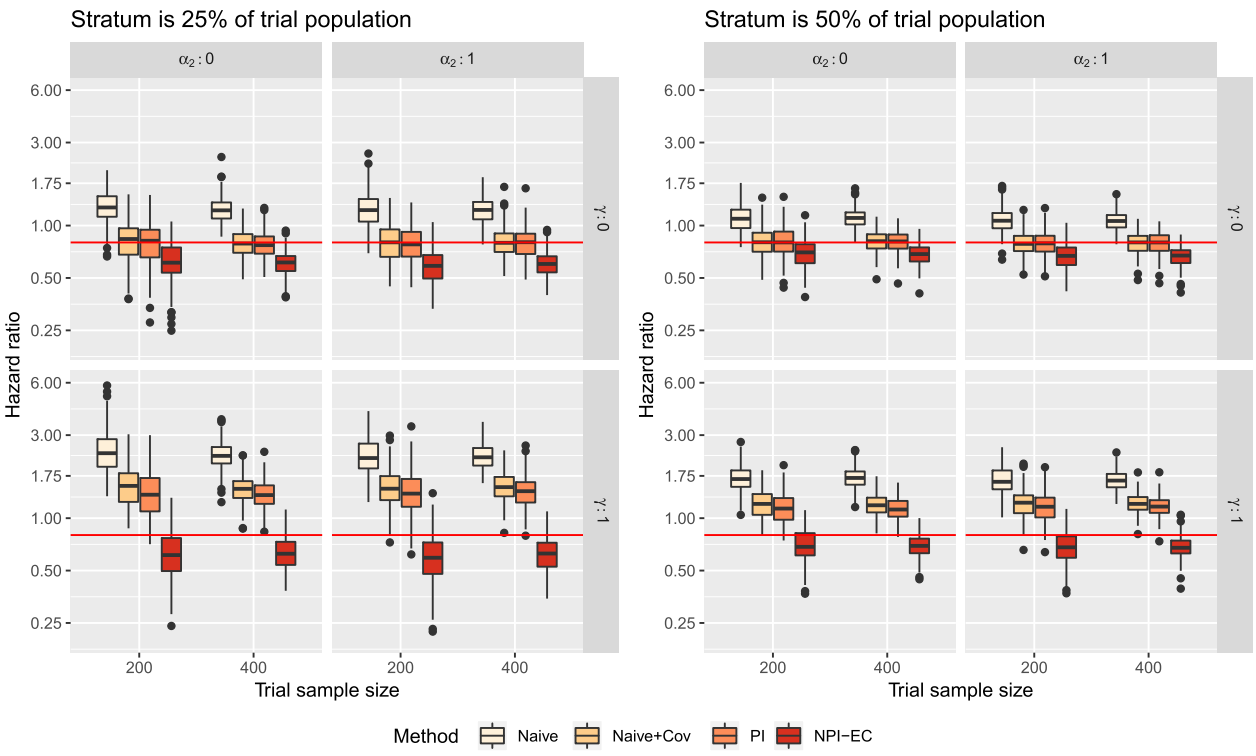

FIGURE 2 Boxplots of log(HR) estimates using naive subpopulation analysis (Naive and Naive + Cov) and joint imputation models (PI and NPI-EC). The horizontal lines indicate the true parameter value used to simulate the data

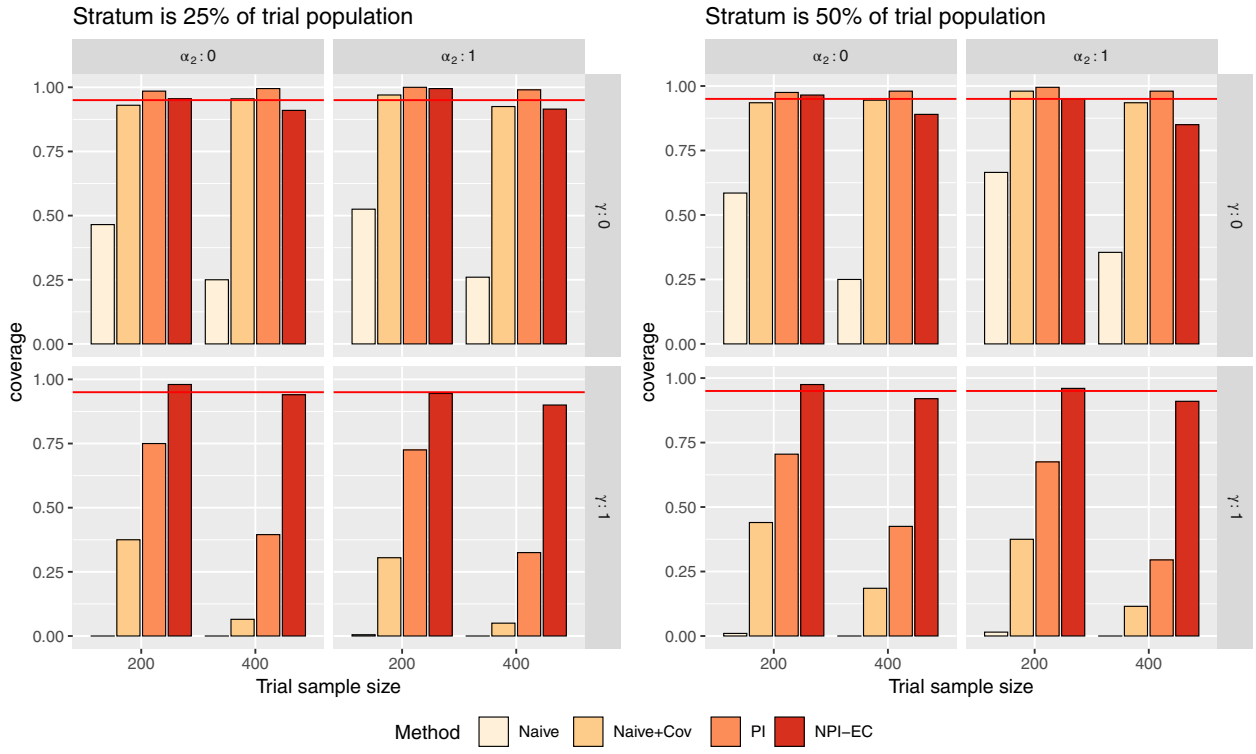

FIGURE 3 Barplots of the coverage of 95% confidence intervals for log(HR) using naive subpopulation analysis (Naive and Naive + Cov) and joint imputation models (PI and NPI-EC). The horizontal lines indicate the nominal coverage of 95%

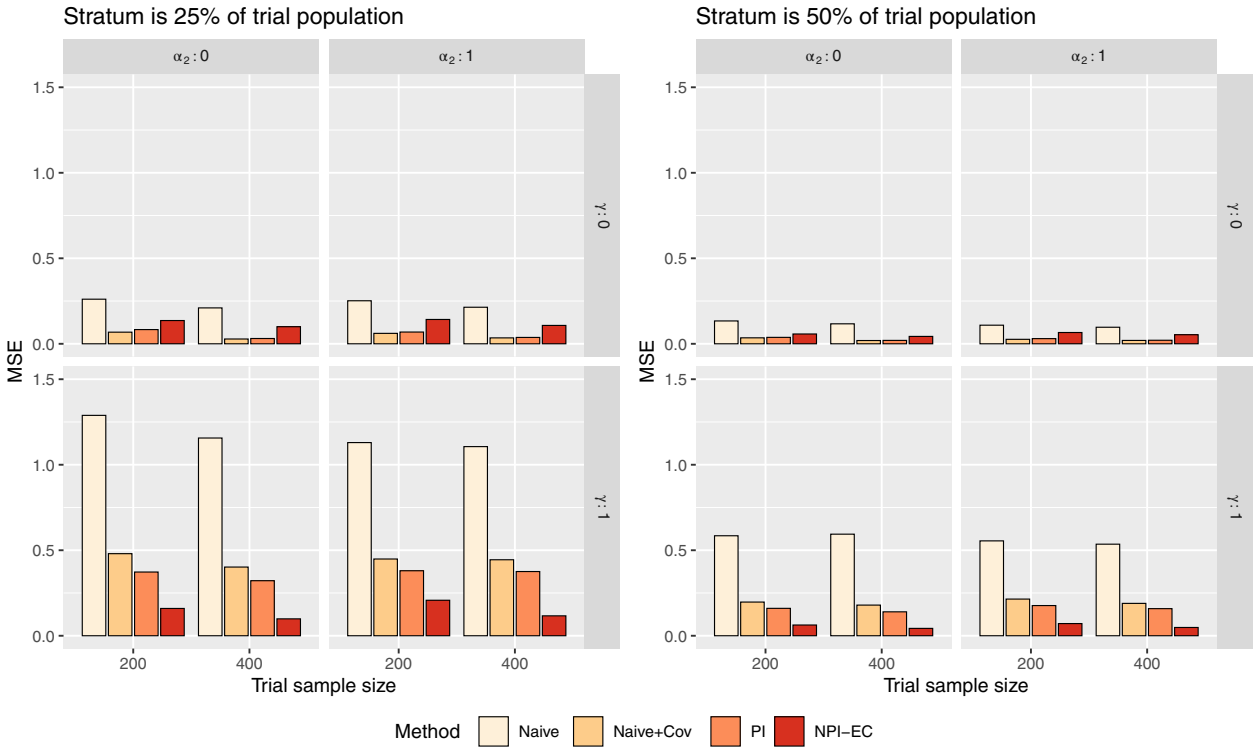

FIGURE 4 Barplots of mean squared errors (MSE) for log(HR) using naive subpopulation analysis (Naive and Naive + Cov) and joint imputation models (PI and NPI-EC)

When the principal ignorability assumption does not hold ($\gamma = 1$), the NPI-EC model performance remained similar to that when $\gamma = 0$ while naive and PI models showed a notable increase in bias and MSE as well as a decrease in coverage. The NPI-EC model slightly underestimates log(HR), as expected, since the dependency between $Y(1)$ and $S(1)$ is stronger than between $Y(0)$ and $S(1)$ ($\eta > \gamma$). Using the estimated sensitivity parameters from the treatment arm is therefore always considered to be an extreme case scenario. The coverage of 95% CI from the NPI-EC model is close to the nominal coverage compared to the other models.

The performance of all models improved with increased sample size (from 200 to 400) and/or principal stratum size (from 25% to 50%), as larger sample size and more events reduce variability. There is a minimal change of model performance for PI and NPI-EC approaches when α_2 changes from 0 to 1, hence illustrating that good prediction of $S(1)$ is not necessary as long as the confounders (here just X_1) are taken into account.

5 | EXAMPLE REVISITED

As a case study, we revisit the phase III oncology trial from Section 2 that compared a new TRT plus SOC versus PBO plus SOC. For consistency with the trial's primary analysis, we use Cox proportional hazard model and use hazard ratio as the effect measure. We are interested in assessing treatment effect in the principal strata of patients who discontinued TRT due to AEs. Table 2 provides the key definitions of this supplemental estimand comparing against the primary estimand. We use the principal ignorability assumption during the estimation of the supplemental estimand. Its sensitivity analysis has consistent attributes with the supplemental estimand in Table 2 except the principal ignorability assumption is modified.

For this purpose, we generated a synthetic dataset based on the real data using the synthpop R package[35] to maintain the confidentiality of the real data while keeping its information structure. Synthesis was performed by syn() function in the order of treatment arm, then prognostic factors conditional on the treatment arm and finally PFS event time conditional on all the information. In the synthetic trial, a high proportion of patients discontinued TRT due to AE in the treatment arm (24%). The joint imputation model is used to evaluate whether such principal stratum of patients who discontinued TRT due to AE still benefited from the treatment overall.

The PFS outcome for the simulated synthetic data for each arm is visualized in a Kaplan–Meier plot in Figure 5, additionally with principal strata of patients who discontinued TRT due to AE ($S(1) = 1$) and those who did not ($S(1) = 0$). Using the synthetic data for the overall study population, a PFS benefit was still observed in favor of the treatment arm (un-adjusted HR = 0.63 with 95% CI: 0.49, 0.83). However, when looking at the patients who discontinued TRT due to AE in the treatment arm, their Kaplan–Meier curve crosses the curve for the control arm after approximately 10 months which raises questions about the magnitude of the treatment effect in this strata of patients.

TABLE 2 Key attributes of the primary and the supplemental estimand

	Primary estimand	Supplemental estimand
Population	Intention-to-treat trial population	Patients who would discontinue TRT due to AEs if they were randomized to treatment arm
Treatment	Randomized study treatment (TRT plus SOC vs. PBO plus SOC) plus any next-line anti-cancer therapies as needed	Randomized study treatment (TRT plus SOC vs. PBO plus SOC) plus any next-line anti-cancer therapies as needed
Variable	Progression free survival	Progression free survival
Intercurrent events	Start of new anti-cancer therapy before disease progression (treatment policy strategy) discontinuation of TRT/PBO and/or SOC due to any reason (treatment policy strategy)	Start of new anti-cancer therapy before disease progression (treatment policy strategy) discontinuation of TRT/PBO and/or SOC due to any reason (treatment policy strategy)
Summary	Hazard ratio and its 95% CI from Cox proportional hazard model	Hazard ratio and its 95% CI from Cox proportional hazard model

Abbreviations: AE, adverse event; CI, confidence interval; PBO, placebo; SOC, standard of care; TRT, treatment.

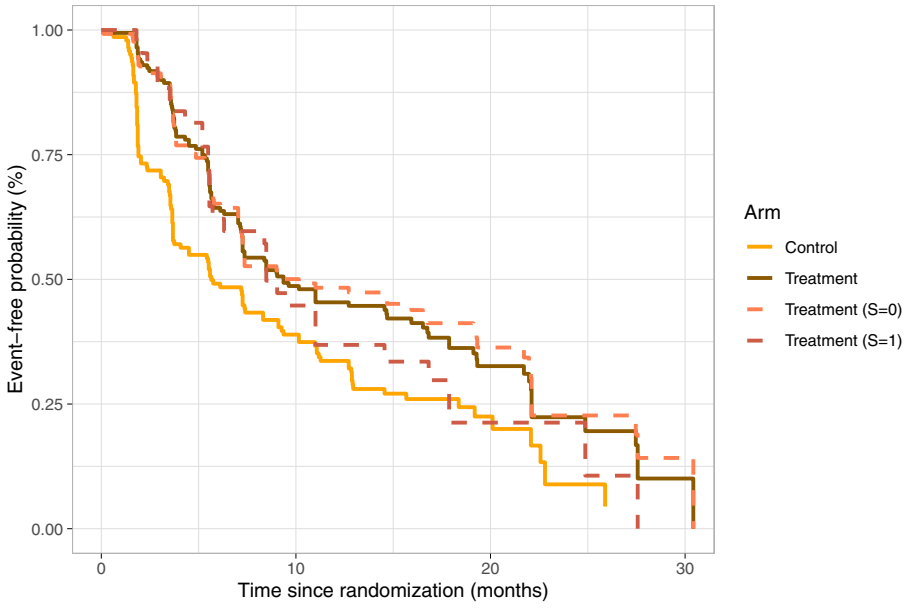

FIGURE 5 Kaplan–Meier plot of PFS for treatment and control arm patients, with separate dashed lines for those who did ($S = 1$) and did not ($S = 0$) discontinue investigational drug due to AE in the treatment arm

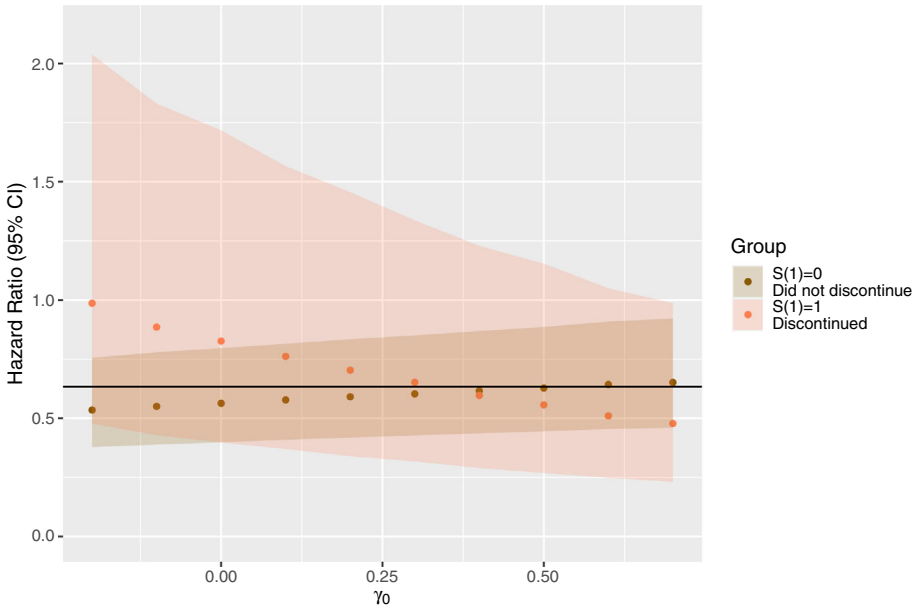

FIGURE 6 Hazard ratios of two principal strata evaluated at different sensitivity parameter values. The shaded area indicates 95% CI. The black horizontal line indicates the overall HR.

For this analysis, we employ a Weibull model with covariates adjustment. Similar to the simulation study, we use JAGS with 15,000 MCMC samples, 5000 samples as run-in and set thinning factor to 5. The baseline covariates used in the joint imputation model were chosen using stepwise selection based on a subset of known clinically relevant prognostic variables and variables that predict treatment discontinuation due to AE. The selected covariates consist of both continuous and categorical variables (covariates-adjusted HR = 0.63 with 95% CI: 0.48, 0.82 for the overall population).

In addition, the parameters are set to have reasonable initial values based on maximum likelihood estimates to facilitate the convergence of MCMC. As discussed in Section 3.3.1, we vary the sensitivity parameter γ in the joint imputation model, within a range that covers the estimated association from the treatment arm as an extreme case and the value 0 as the case when principal ignorability holds.

The treatment effect is summarized using hazard ratio for each MCMC iteration, using a Cox proportional hazard model based on patients with $S(1) = 1$, then combined using Rubin's rule. The final hazard ratios and their 95% CI are visualized for both the principal stratum of interest and the complement stratum in Figure 6, that is, patients who would discontinue TRT due to AE and those who would not.

The estimated $\widehat{\eta}$ is 0.26 (95% CI: −0.16, 0.68), indicating a weak association between $S(1)$ and $Y(1)$ conditional on the covariates. This implies patients who discontinued TRT due to AE on the treatment arm have higher hazards for PFS event. Hazard ratios of patients belonging to the principal stratum is shown in blue and that for complement stratum is shown in red. The confidence interval is wider in the $S(1) = 0$ group since the number of patients who discontinued TRT is only about 24% of the treatment arm from the synthetic data. Nevertheless, a consistent PFS treatment benefit (HR < 1) is observed across the range of γ_0 shown in Figure 6. Note the range of γ_0 covers 95% CI of the $\widehat{\eta}$ estimate. If principal ignorability indeed holds ($\gamma_0 = 0$), the expected HR for the principal stratum of $S(1) = 1$ is 0.82 (95% CI: 0.40, 1.69), which is higher than the naive HR from synthetic data (un-adjusted HR of 0.75 with 95% CI: [0.50, 1.12] and covariates-adjusted HR of 0.76 with 95% CI: [0.51 1.14]).

When $\gamma_0 < 0$, $S(1) = 1$ lowers the hazard of PFS on the control arm hence the treatment benefit is smaller. The HR is around 1.0 at $\gamma_0 = -0.16$, indicating the treatment effect is no longer positive if γ_0 decreases beyond this point. As γ_0 increases, the effect of $S(1) = 1$ on $Y(0)$ shifts from prolonging PFS to shortening PFS according to Equation (5). Since information flows both directions in the joint imputation model, for a large positive γ_0, patients with shorter observed $Y(0)$ are more likely to get imputed $S(1) = 1$. Hence the HR in $S(1) = 1$ stratum decreases as γ_0 increases, while for the complement stratum HR increases with γ_0. The R code used for conducting this analysis as well as a more flexible Weibull mixture model implementation are available in the Appendix as a R markdown file.

6 | SUMMARY AND DISCUSSIONS

We have developed a framework of sensitivity analysis to stress test the assumption of principal ignorability. This assumption is utilized in the context of principal stratum estimands, where outcome and principal stratum membership are assumed to be independent given a set of baseline covariates X. This sensitivity analysis uses a joint imputation model on the principal stratum membership in the treatment arm $S(1)|X$ and a time-to-event outcome in the control arm $Y(0)|S(1), X$. A sensitivity parameter γ induces an association between $Y(0)$ and $S(1)$, with $\gamma = 0$ being when the principal ignorability holds. One can select a range of plausible values for γ to assess how the treatment effect in the stratum of interest changes with different levels of violation of the principal ignorability assumption.

The sensitivity analysis can be used in various clinical research settings. When main results are reported in a randomized trial (often based on utilizing the treatment policy strategy for treatment discontinuation or switches), it may be of interest to evaluate the treatment effect of a drug in a certain subpopulation. If this subpopulation is defined by intercurrent events instead of baseline characteristics, then the principal stratum approach can be considered for handling such subpopulation of interest. Based on the simulation study, it may suffice to use a naive approach with covariates adjustment for estimation under principal ignorability assumption. However, in many cases, the principal ignorability assumption is not verifiable. Hence, it can be misleading to report a single summary measure for the treatment effect in a principal stratum. Instead, a range of treatment effects corresponding to different sensitivity parameter values should be considered as supportive analysis. Analogous to a tipping point analysis,[36] the sensitivity analysis can be used to identify tipping points where the treatment effect in the stratum of interest may no longer be positive or clinically relevant.

The sensitivity parameter represents the association between principal stratum membership in treatment arm and clinical outcome in control arm, yet these two variables do not co-exist in the 'same world' for the same patient, which in practice poses a challenge in its interpretation. However, we can use η, the association between principal stratum membership and outcome in treatment arm, as a reference level to help with the interpretation. This is precisely what we used in the simulation study as part of the NPI-EC method. By assuming the cross-world association γ being always weaker than the same world association η, a clinical interpretation can be first attributed to a value of η and then consider the same value of γ as an extreme case setting.

Mattei et al.[37] investigated a situation of treatment switching, that has many characteristics similar to the situations considered here. They utilize a fully Bayesian inference approach for the switching problem. While it is desirable to use the same analysis model as used for the primary analysis, it would be also interesting to extend their approach for the situation considered here. Jiang et al.[38] discussed the impact of model misspecifications in $S(1)|X$ and $Y(0)|S(1), X$ on

the estimators, which are also relevant for our approach. In the presented approach, the time to treatment discontinuation $T(1)$ itself was not modeled directly (only $S(1)$, in essence a dichotomized version of $T(1)$ was modeled). It is an interesting area of further work to extend the joint multiple imputation approach towards handling time to treatment discontinuation as well. In doing so, it will offer the possibility to modify the treatment effect of interest from the standard PFS definition as treatment effect after randomization to investigate in particular the treatment effect after treatment discontinuation.

Our proposed approach complements existing principal stratum methods for time-to-event endpoints. The sensitivity analysis with the option of identifying tipping points can be considered as a valuable alternative that offers more appropriate results than a single summary measure under unverifiable assumptions.

ACKNOWLEDGMENTS

The authors would like to thank Alessandra Mattei, Mathilde Kaper, and Arunava Chakravartty for their discussions on the topic and providing valuable comments on early versions of the manuscript.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

Research data are not shared; the synthetic data and its generating process used are described in the manuscript and the Supporting information appendix.

ORCID

Craig Wang https://orcid.org/0000-0003-1804-2463
Björn Bornkamp https://orcid.org/0000-0002-6294-8185

REFERENCES

1. ICH. Addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials E9(R1) 2019.
2. Lipkovich I, Ratitch B, Mallinckrodt CH. Causal inference and estimands in clinical trials. *Stat Biopharm Res*. 2020;1:54-67.
3. Frangakis Constantine E, Rubin DB. Principal stratification in causal inference. *Biometrics*. 2002;58:21-29.
4. Björn B, Kaspar R, Jianchang L, et al. Principal stratum strategy: potential role in drug development. *Pharm Stat*. 2021;20:737-751.
5. Hernán Miguel A, Daniel S. Cautions as regulators move to end exclusive reliance on intention to treat. *Ann Intern Med*. 2018;168:515-516.
6. Scharfstein DO. A constructive critique of the draft ICH E9 addendum. *Clin Trials*. 2019;16:375-380.
7. Judea P. Principal stratification–a goal or a tool? *Int J Biostat*. 2011;7:20.
8. Mealli F, Mattei A. A refreshing account of principal stratification. *Int J Biostat*. 2012;1:1-37.
9. Yiyue L, Jones Michael P, Wanjie S. Estimation of causal effects in clinical endpoint bioequivalence studies in the presence of intercurrent events: noncompliance and missing data. *J Biopharm Stat*. 2019;29:151-173.
10. Yasutaka Chiba, Van der Weele Tyler J A simple method for principal strata effects when the outcome has been truncated due to death *Am J Epidemiol* 2011;173:745–751.
11. Amy R, Hudgens Michael G, Gilbert Peter B, Fine Jason P. Nonparametric bounds and sensitivity analysis of treatment effects. *Stat Sci*. 2014;29:596.
12. Booil J, Stuart Elizabeth A. On the use of propensity scores in principal causal effect estimation. *Stat Med*. 2009;28:2857-2875.
13. Peng D, Jiannan L. Principal stratification analysis using principal scores. *J R Stat Soc: Series B (Stat Methodol)*. 2017;79:757-777.
14. Avi F, Fabrizia M, Luke M. Principal score methods: assumptions, extensions, and practical considerations. *J Educ Behav Stat*. 2017;42:726-758.
15. Zhichao J, Peng D. Identification of causal effects within principal strata using auxiliary variables. *Stat Sci*. 2021;36:493-508.
16. Fabrizia M, Barbara P. Using secondary outcomes to sharpen inference in randomized experiments with noncompliance. *J Am Stat Assoc*. 2013;108:1120-1131.
17. Daniel W, Edwards Jessie K, Cole Stephen R, Platt Robert W, Mumford Sunni L, Schisterman Enrique F. Imputation approaches for potential outcomes in causal inference. *Int J Epidemiol*. 2015;44:1731-1737.
18. Peng D, Fan L. Causal inference: a missing data perspective. *Stat Sci*. 2018;33:214-237.
19. Jerome C, William H, Cuyler HE, Lilienfeld Abraham M, Shimkin Michael B, Wynder Ernst L. Smoking and lung cancer: recent evidence and a discussion of some questions. *J Natl Cancer Inst*. 1959;22:173-203.
20. Rosenbaum PR, Rubin DB. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J R Stat Soc: Series B (Methodol)*. 1983;45:212-218.

21. Peng D, Van der Weele Tyler J. Sensitivity analysis without assumptions. *Epidemiol*. 2016;27:368-377.
22. Stuart Elizabeth A, Booil J. Assessing the sensitivity of methods for estimating principal causal effects. *Stat Methods Med Res*. 2015;24: 657-674.
23. Alessandra M, Laura F, Fabrizia M. Assessing principal causal effects using principal score methods. In: Zubizarreta José R, Stuart Elizabeth A, Small Dylan S, Rosenbaum Paul R, eds. *Handbook of Matching and Weighting Adjustments for Causal Inference*; 2022 Forthcoming.
24. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer*. 2009;45:228-247.
25. Evgeny D, Yiyun Z, Kapildeb S, et al. Estimands and the patient journey: addressing the right question in oncology clinical trials. *JCO Precis Oncol*. 2019;3:1-10.
26. Jerzy S-N, Dabrowska DM, Speed TP. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Stat Sci*. 1990;5:465-472.
27. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*. 1974;66:688-701.
28. Björn B, Georgina B. 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 Biopharm Res*. 2020;12:19-28.
29. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons; 1987.
30. Torben M, Stijn V, Kragh AP. Subtleties in the interpretation of hazard contrasts. *Lifetime Data Anal*. 2020;26:833-855.
31. Liao Jason JZ, Frank LG. A flexible parametric survival model for fitting time to event data in clinical trials. *Pharm Stat*. 2019;18: 555-567.
32. McCandless Lawrence C, Paul G. A comparison of Bayesian and Monte Carlo sensitivity analysis for unmeasured confounding. *Stat Med*. 2017;36:2887-2901.
33. Carvalho CM, Polson NG, Scott JG. The horseshoe estimator for sparse signals. *Biometrika*. 2010;97:465-480.
34. Martyn Plummer. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling in *Proceedings of the 3rd international workshop on distributed statistical computing*;124:1–10 2003.
35. Beata N, Raab GM, Chris D. Synthpop: bespoke creation of synthetic data in R. *J Stat Softw*. 2016;74:1-26.
36. Ilya L, Bohdana R, Michael O'K. Sensitivity to censored-at-random assumption in the analysis of time-to-event endpoints. *Pharm Stat*. 2016;15:216-229.
37. Alessandra Mattei, Fabrizia Mealli, Peng Ding. Assessing causal effects in the presence of treatment switching through principal stratification 2020. arXiv.
38. Zhichao Jiang, Yang Shu, Peng Ding. Multiply robust estimation of causal effects under principal ignorability 2020. arXiv.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Wang C, Zhang Y, Mealli F, Bornkamp B. Sensitivity analyses for the principal ignorability assumption using multiple imputation. *Pharmaceutical Statistics*. 2023;22(1):64-78. doi:10.1002/pst.2260