Surrogacy validation for time-to-event outcomes with illness-death frailty models

Emily K. Roberts 1 | Michael R. Elliott 2,3 | Jeremy M. G. Taylor 2

1 Department of Biostatistics, University of Iowa, Iowa City, Iowa, USA
2 Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, USA
3 Survey Methodology Program, Institute for Social Research Ann Arbor, Ann Arbor, Michigan, USA

Correspondence
Emily K. Roberts, Department of Biostatistics, University of Iowa, 145 N. Riverside Dr, Iowa City, IA 52242, USA. Email: emily-roberts-1@uiowa.edu

Funding information
National Institutes of Health, Grant/Award Number: CA129102

Abstract
A common practice in clinical trials is to evaluate a treatment effect on an intermediate outcome when the true outcome of interest would be difficult or costly to measure. We consider how to validate intermediate outcomes in a causally-valid way when the trial outcomes are time-to-event. Using counterfactual outcomes, those that would be observed if the counterfactual treatment had been given, the causal association paradigm assesses the relationship of the treatment effect on the surrogate outcome with the treatment effect on the true, primary outcome. In particular, we propose illness-death models to accommodate the censored and semicompeting risk structure of survival data. The proposed causal version of these models involves estimable and counterfactual frailty terms. Via these multistate models, we characterize what a valid surrogate would look like using a causal effect predictiveness plot. We evaluate the estimation properties of a Bayesian method using Markov chain Monte Carlo and assess the sensitivity of our model assumptions. Our motivating data source is a localized prostate cancer clinical trial where the two survival outcomes are time to distant metastasis and time to death.

Keywords
Bayesian methods, clinical trial, illness-death model, surrogacy validation, time-to-event data

1 | INTRODUCTION

Biomarkers such as HbA1c in diabetes trials, CD4 counts in AIDS studies, or immune responses in vaccine development are all potential targets on which to measure the effect of a given treatment in a clinical trial. In fact, these outcomes can be assessed at some intermediate point of the trial that occurs substantially before long-term, primary outcomes such as diabetes vascular complications, disease progression, or infection of a virus, respectively, may be measured. In oncology trials where time-to-event endpoints are common, it can take many years to accrue enough observed events, such as death from any cause, to complete the study (Kemp & Prasad, 2017). An appropriate surrogate endpoint can be measured earlier to serve as a substitute indicator for if a treatment effect exists on some true clinical outcome of interest. Broadly, we would like to make an assessment about the treatment based on the information that is available on the intermediate outcome.
in the trial at a relatively early time after randomization $\tau_S$, instead of basing an assessment about the treatment only on the information about the primary outcome at a significantly later time point $\tau_T$.

During the course of cancer treatment, patients are typically followed over the study period while several time-to-events are measured such as local, regional, and distant failure as the cancer progresses in stages. In this work, our data come from a prostate cancer clinical trial assessing the addition of an antiandrogen therapy to an existing regimen (Shipley et al., 2017). Here, the primary trial outcome is death from any cause, and the intermediate occurrence of distant metastases is also of interest. For these patients, death from prostate cancer will only occur if the person has had metastases. However, some men will experience death from different causes during follow-up with or without experiencing distant metastases spreading first. Overall survival (OS) is therefore a mixture of two death types, death from prostate cancer and death from other causes. Mechanistically understanding whether the time-to-event outcome of distant metastases is a desirable surrogate for OS in this setting may be beneficial for clinicians and trialists.

### 1.1 Current limitations for validation with time-to-event data

If the intermediate outcome is determined to be a valid surrogate and $\tau_S$ is substantially earlier than $\tau_T$, then the intermediate outcome could be very useful in future trials. Given the substantial risk of potentially using an invalid surrogate outcome in a large-scale trial, rigorous standards have been proposed to validate a surrogate (VanderWeele, 2013). The first criteria to determine the validity of candidate surrogate outcomes were suggested by Prentice (1989) which test whether a treatment affects the true outcome only through the pathway of the surrogate outcome. While the criteria are applicable to different outcomes such as time-to-event outcomes that we will be focusing on, they involve regression models that rely on conditioning on the observed value of the surrogate outcome, leading to a noncausal interpretation. More recent frameworks to determine if a surrogate marker is appropriate for use in a future trial can be broadly grouped into the causal effects and causal association paradigms (Joffe & Greene, 2009). The causal association framework aims to evaluate the relationship of the treatment effect on the surrogate $S$ with the treatment effect on the true clinical outcome $T$. These methods are often built upon counterfactual outcomes $T(z)$, which are the clinical outcomes of interest, and $S(z)$, the surrogate outcomes, where the notation $Z = z$ represents treatment under either the observed or counterfactual assignment.

Methods within the causal association framework have been proposed for trials where the true outcome $T$ is a time-to-event outcome for non-time-to-event surrogate outcomes. Tanaka et al. (2017) and Gao (2012) both consider a binary surrogate for a survival primary outcome using principal stratification methods (Frangakis & Rubin, 2002). Taylor et al. (2015) propose a Gaussian copula model with a survival outcome for $T$ and ordinal outcome $S$, and others allow for a continuous $S$ and time-to-event $T$ (Gabriel & Gilbert, 2014; Gabriel et al., 2015; Qin et al., 2008). Causal solutions for validation become more challenging when the surrogate is also subject to censoring. Instead, others such as Parast et al. (2017) rely on measures such as proportion explained, and Hsu et al. (2015); Vandenberghe et al. (2018), and Weir et al. (2021) address time-varying surrogates using mediation approaches and proportion mediated metrics within the causal effects paradigm.

To our knowledge, the setting where both the surrogate and primary outcome are time-to-event outcomes has not been fully addressed within the principal stratification framework. Building on the work of Frangakis and Rubin (2002), we aim to develop a corresponding causal effect predictiveness (CEP) curve proposed by Gilbert and Hudgens (2008) to validate a surrogate outcome when both the surrogate and primary outcome are time-to-event. The key to obtaining a causal assessment in this paradigm is classifying individuals based on their set of potential values of the posttreatment variable, which is the surrogate outcome in our setting. In a simple case where $S$ and $T$ are Gaussian outcomes and $Z$ takes on the value 0 or 1, the analog to surrogate-specific strata and the corresponding CEP curve for validation is based on the quantity $E(T(1) - T(0)|S(1) - S(0) = s)$. Notably this is for when $S$ and $T$ are observed, so it needs adaptation when $S$ and $T$ may be censored. The CEP curve involves two contrasts, which we will denote by $\Delta T$ and $\Delta S$, and is based on the conditional distribution of $\Delta T$ given $\Delta S = s$ as a function of $s$. Briefly, the CEP criteria intuitively assert that there be no average treatment effect on the primary outcome for the strata of patients defined by no treatment effect on the surrogate outcome (i.e., $s = 0$), and conversely that there exist an overall treatment effect on the primary outcome for the strata of patients defined by a treatment effect on the surrogate outcome (i.e., $s \neq 0$). A comparable contrast and consideration of principal strata when $T(z)$ and $S(z)$ are subject to censoring and a semicompeting risk structure will be explored in this paper.

Outside of the surrogacy validation setting, semicompeting risks based on counterfactual hazards have been explored (Huang, 2021). Within the principal stratification framework, unobserved outcomes due to truncation by death can be
addressed by defining strata based on survivorship cohorts; causal effects are defined for individuals who would survive regardless of the assigned treatment (Comment et al., 2019; Xu et al., 2020; Zhang & Rubin, 2003). Models do exist that are suitable for these data by accounting for semicompeting risks, though are there complications in defining proper estimands (see Bühler et al., 2022). Specifically, whereas the previous CEP validation quantities suggest conditioning on counterfactual surrogate outcomes, in our setting the surrogate outcome may not be well-defined if it is not observed before the primary outcome; it may not be possible to condition on strata defined by $S(0)$ and $S(1)$ occurring or not by time $\tau$. For example, while we can construct

$$P(T(1) < \tau \mid S(0) \geq \tau, S(1) \geq \tau)$$

or

$$P(T(1) < \tau \mid T(0) \geq \tau, S(0) \geq \tau, S(1) \geq \tau)$$

for some time $\tau$, it is not clear if either would be a principled estimand to use for validation with our outcome types.

The outline of our approach is to specify an illness-death model for counterfactual outcomes $S(0), S(1), T(0), T(1)$. Then using the principal stratification approach, we define a metric that will be used to assess whether $S$ is a valid surrogate for $T$, this metric being the CEP curve. The CEP curves involve two contrasts, $\Delta T$ and $\Delta S$. In Section 2, we describe our causal modeling approach that incorporates the semicompeting risk structure of the data using an illness-death model and counterfactual hazards. Similar to previous work for repeated measurements with subject-specific random effects (Roberts et al., 2023), we utilize frailty terms to capture subject-specific heterogeneity and allow dependence among the transitions of the illness-death model. The CEP curve is shown to be a function of the parameters in the illness-death model. This approach is novel in our single trial setting with subject-level, counterfactual outcomes (in contrast to joint frailty-copula models used for meta-analysis in Emura et al. (2017) and Sofeu et al. (2020)). Section 3 describes our proposed CEP quantities and explores what an ideal surrogate would look like. In Section 4, we provide the likelihood of the illness-death model and propose Bayesian methods to estimate these parameters and hence to estimate the CEP curve. A simulation study is provided in Section 5 with a real data analysis from a prostate cancer trial in Section 6. Discussion and future work are provided in Section 7.

2 | ILLNESS-DEATH APPROACH

The structure of the illness-death model is a natural way to describe data with the semicompeting risk structure and has potential use for surrogacy validation (O’Quigley & Flandre, 2012). Here, we consider counterfactual illness-death models and the principal stratification framework. Let $T_{jk}(z)$ denote the gap time between two states ($j = 1, 2, k = 2, 3$) and corresponding transition intensities $\lambda_{jk}^z$ between states in the treatment-specific illness-death models for treatment $Z = z$ as shown in Figure 1. The model has some connection to the models in the Prentice criteria, as shown in the Supporting Information, but the Prentice criteria models do not involve counterfactual quantities.

We propose a more rigorous and flexible strategy to identify a consistent surrogate using potential outcomes and counterfactual illness-death models in pursuit of a causal interpretation (VanderWeele, 2013). In the model, we propose and explore in detail in the following sections, each counterfactual arm has its own set of transition hazard models. We will first consider all counterfactual quantities that appear in the complete data likelihood for the proposed model.

2.1 | Defining causal quantities based on hazards and frailty models

We propose to model the transition hazards that correspond to the gap times $T_{jk}(z)$ between states $\tilde{S}(z)$ and $\tilde{T}(z)$ in Figure 1 (once an individual in treatment arm $z$ experiences the surrogate outcome, they move into state $\tilde{S}(z)$ for example). Shared or common frailty terms, which quantify the dependence between the different processes within the same person, can
provide information on the dependence structure between the time to intermediate event and the time to terminating event in multistate models (Xu et al., 2010; Zhang et al., 2014). In models for time-to-event data, frailties are commonly incorporated to model correlation among events, to allow for heterogeneity among individuals, or to capture the effect of some omitted covariate. In our setting, we consider both counterfactual outcomes and transitions, and we want to allow for possible dependence between the counterfactual outcomes. As this association is integral to the value of the surrogate, we propose to use illness-death frailty models where the hazards are linked via frailty terms. Here, we consider multiple hazards with frailties both to allow dependence across state transitions and to link observable transitions in arm \( Z = z \) to the counterfactual transitions for \( Z = 1 - z \).

For a single time-to-event and a general frailty \( \omega \), the hazard can be written \( \lambda(t|X, \beta, \omega, \kappa) = \lambda_0(t) \exp(\kappa \omega + X\beta) \), where \( X \) is a baseline covariate and \( \omega \) has some prespecified distribution and may have an associated coefficient parameter \( \kappa \). Various assumptions can be made about the frailty term \( \omega \), such as that it follows a Normal or Gamma distribution, for simplicity and computational feasibility. For the illness-death models specified in Figure 1, a set of the six correlated frailties are required, one for each model. However, for identifiability and computational concerns, we impose some restrictions and simplifying assumptions. We initially propose two different formulations of the sets of models, and we exclude consideration of baseline covariates \( X \).

### 2.1.1 Model A using time-dependent covariates

For \( z = 0 \),

\[
\lambda^0_{12}(t|\omega^0_{12}) = \lambda^0_{12,0}(t) \exp(\kappa^0_0 \omega^0_{12})
\]

\[
\lambda^0_{13}(t|\omega^0_{13}) = \lambda^0_{13,0}(t) \exp(\kappa^0_1 \omega^0_{13})
\]

\[
\lambda^0_{23}(t|T^0_{12}(0), \omega^0_{23}) = \lambda^0_{23,0}(t - T^0_{12}(0)) \exp(\kappa^0_2 \omega^0_{23} + \theta^0_{23} T^0_{12}(0))I(t > T^0_{12}(0)).
\]

Similarly for \( z = 1 \),

\[
\lambda^1_{12}(t|\omega^1_{12}) = \lambda^1_{12,0}(t) \exp(\kappa^1_0 \omega^1_{12})
\]

\[
\lambda^1_{13}(t|\omega^1_{13}) = \lambda^1_{13,0}(t) \exp(\kappa^1_1 \omega^1_{13})
\]

\[
\lambda^1_{23}(t|T^1_{12}(1), \omega^1_{23}) = \lambda^1_{23,0}(t - T^1_{12}(1)) \exp(\kappa^1_2 \omega^1_{23} + \theta^1_{23} T^1_{12}(1))I(t > T^1_{12}(1)),
\]

where \( T_{jk} \) represent gap times and \( t \) represents time from randomization. Therefore, \( T_{12i} \) as a covariate is the time that subject \( i \) moves into state \( \tilde{S} \), or in other words experiences the surrogate outcome, and we model the third transition using a clock reset for \( \tilde{\tau}_{23} \) (i.e., the time scale is time since the surrogate outcome occurred, \( t - T_{12}(z) \)). We include \( \theta^1_{23} \) in the \( \chi_{13} \) model as the coefficient for our time-dependent covariate \( T_{12} \). The purpose is to capture the effect of this transition time, and the time that an individual experiences the surrogate outcome may help to assess the strength of association between \( S \) and \( T \).

The restrictions and assumptions we will be considering are to make \( \omega^0_{131} = \omega^0_{231} \) and to set some of the \( \kappa^0_{jk} = 1 \). If the \( \kappa \) parameters vary, they essentially influence how variable the frailty terms are. We will refer to \( \kappa \) as frailty coefficients. One rationale for assuming \( \omega^0_{131} = \omega^0_{231} \) in this setting is that both are frailties that influence time to death from other causes in our motivating trial. For example, since our primary outcome variable is death from any cause, we may expect that some men will die of old age. It may be reasonable to expect that an individual may have their own propensity for experiencing death from other causes irrespective of whether or not the surrogate outcome has occurred. Another consideration is that by including the coefficient for our time-varying covariate, \( \theta^1_{23} \), the model captures the magnitude of the effect for the time it takes to experience the intermediate outcome. This makes it more plausible that certain frailties are equal and conditional independence assumptions may be more likely. Finally, the frailties capture heterogeneity on the individual
level. There may still be heterogeneity on the population level for the variability in the hazard of going from baseline to the primary outcome directly or from the surrogate to primary outcome which can be reflected in the baseline hazards. We explore these variations in later sections.

2.1.2 Model B using multiple frailties in place of time-dependent covariates

We include an alternate option to incorporate the dependence between the different transitions such as a model that includes two frailty terms in the $\tilde{S} \rightarrow T$ transition

$$
\lambda_0^{12}(t|\omega_{12}^0) = \lambda_0^{12,0}(t) \exp(\chi_{12}^0 \omega_{12}^0)
$$

$$
\lambda_0^{13}(t|\omega_{13}^0) = \lambda_0^{13,0}(t) \exp(\chi_{13}^0 \omega_{13}^0)
$$

$$
\lambda_0^{23}(t|T_{12}(0),\omega_{13}^0,\omega_{12}^0) = \lambda_0^{23,0}(t-T_{12}(0)) \exp(\chi_{12}^0 \omega_{12}^0 + \chi_{13}^0 \omega_{13}^0) I(t > T_{12}(0))
$$

$$
\lambda_1^{12}(t|\omega_{12}^1) = \lambda_1^{12,0}(t) \exp(\chi_{12}^1 \omega_{12}^1)
$$

$$
\lambda_1^{13}(t|\omega_{13}^1) = \lambda_1^{13,0}(t) \exp(\chi_{13}^1 \omega_{13}^1)
$$

$$
\lambda_1^{23}(t|T_{12}(1),\omega_{13}^1,\omega_{12}^1) = \lambda_1^{23,0}(t-T_{12}(1)) \exp(\chi_{12}^1 \omega_{12}^1 + \chi_{13}^1 \omega_{13}^1) I(t > T_{12}(1)).
$$

The motivation of this model is an alternative way to capture the subject-specific relationship between the different transitions via the $\chi_{12}^z$ and $\chi_{13}^z$ coefficients. This model does not include $T_{12}$ as a time-varying covariate. When we assume $\omega_{23} = \omega_{13}^z$, the key difference between models A and B is the way in which the transition from baseline to the intermediate outcome and the time following that transition are related; these are linked using either a time-varying covariate (in model A) or another frailty term (in model B). Again, the frailty coefficients $\chi$ can be thought of parameters that increase or decrease the magnitude of the effect of the frailties. We would not expect $\chi_{12}^z$ and $\chi_{12}^1$ to be necessarily equal across the models given the different assumptions in each model.

2.1.3 Frailty structures

In its most generality, model A has six correlated frailties, which we assume have a multivariate normal distribution.

$$
\begin{pmatrix}
\omega_{12}^0 \\
\omega_{12}^1 \\
\omega_{13}^0 \\
\omega_{13}^1 \\
\omega_{23}^0 \\
\omega_{23}^1
\end{pmatrix} \sim N
\begin{pmatrix}
0 \\
0 \\
0 \\
0 \\
0 \\
0
\end{pmatrix},
\begin{pmatrix}
1 & \rho_S & \rho_{00} & \rho_{01} & \rho_{S1} & \rho_{S2} \\
\rho_{10} & 1 & \rho_{11} & \rho_{S3} & \rho_{S4} \\
\rho_{T1} & \rho_{T1} & 1 & \rho_{T3} & \rho_{T4} \\
\rho_{T2} & \rho_{T2} & \rho_{T3} & 1 & \rho_{ST} \\
\rho_{ST} & \rho_{ST} & \rho_{ST} & \rho_{ST} & 1
\end{pmatrix}
$$

While this model has a very general form, it may not be necessary or even desirable to consider this level of generality. We will be focusing on special cases of this general model, which we think are appropriate for the setting of surrogacy assessment.

To reduce the number of frailties to estimate to four in model A that we will focus on, we assume that both transition paths to the primary outcome have the same frailty ($\omega_{13}^z = \omega_{23}^z$). As discussed above, since the terminal event is death from any cause, it seems justifiable to assume that conditional on all other terms in the model, frailties toward death from any cause would be the same on the individual level with or without the occurrence of the intermediate event. This
assumption will be useful for estimation since $T_{23}$ is not defined for all individuals. With this assumption, our transition models from $S$ to $T$ in model A can be written

$$
\lambda_{23}^0(t|T_{12}(0), \omega_{12i}^0) = \lambda_{23,0}^0 (t - T_{12}(0)) \exp(\kappa_{23}^0 \omega_{12i}^0 + \theta_{23}^0 T_{12}(0))I(t > T_{12}(0))
$$

$$
\lambda_{23}^1(t|T_{12}(1), \omega_{12i}^1) = \lambda_{23,0}^1 (t - T_{12}(1)) \exp(\kappa_{23}^1 \omega_{12i}^1 + \theta_{23}^1 T_{12}(1))I(t > T_{12}(1)).
$$

Ideally, we could allow $\kappa_{23}$ to take on different values from $\kappa_{13}$ to accommodate different amounts of dependence between the transitions. Then for both models A and B, we consider the joint distribution

$$
\begin{pmatrix}
\omega_{12i}^0 \\
\omega_{12i}^1 \\
\omega_{13i}^0 \\
\omega_{13i}^1
\end{pmatrix}
\sim N
\begin{pmatrix}
0 & \rho_S & \rho_{00} & \rho_{01} \\
0 & 1 & \rho_{01} & \rho_{11} \\
0 & \rho_S & 1 & \rho_T \\
0 & 1 & \rho_T & 1
\end{pmatrix}.
$$

In most of the work presented here, we will also assume $\omega_{12i}^0 \perp \omega_{13i}^0$ (the frailties for an individual are independent across states), meaning $\rho_{00} = \rho_{01} = \rho_{11} = \rho_{10} = 0$. Only $\rho_{00}$ and $\rho_{11}$ of these are identifiable. We thus assume

$$
\begin{pmatrix}
\omega_{12i}^0 \\
\omega_{12i}^1
\end{pmatrix}
\sim N
\begin{pmatrix}
0 & \rho_S \\
0 & 1
\end{pmatrix}
\quad \text{and} \quad
\begin{pmatrix}
\omega_{13i}^0 \\
\omega_{13i}^1
\end{pmatrix}
\sim N
\begin{pmatrix}
0 & \rho_T \\
0 & 1
\end{pmatrix}.
$$

This type of assumption may aid in estimation. We could instead impose a strong assumption of shared frailties for each arm: $\omega_{12i}^0 = \omega_{13i}^0 = \omega_{23i}^0$ and $\omega_{12i}^1 = \omega_{13i}^1 = \omega_{23i}^1$. The motivation for this comes from considering the frailty as representing an omitted covariate. We do not further pursue this assumption.

### 2.2 Identifiability and sensitivity analysis

Certain parameters within our model are nonidentifiable because they describe relationships between counterfactual variables ($\rho_S$ and $\rho_T$ for example), while others are “barely” identifiable (the combination of the baseline hazard, frailties, and the $\kappa$ parameters, for example) and are therefore hard to estimate. In particular, the frailty terms are weakly identified based on which events, the surrogate and/or primary outcome, are actually observed. Since we have made modeling assumptions to aid in estimation, we can evaluate the sensitivity of the assumed models in several ways. Because the parameters $\rho_S$ and $\rho_T$ in the complete data likelihood are not identifiable, they will be fixed at preset values in our proposed method (and later we will discuss if the complete data likelihood is necessary). Based on biological considerations under the counterfactual framework, we may not expect these correlation parameters to be negative or exactly equal to one. Furthermore, we can vary which frailties are assumed to be independent or equal, alter which values of $\kappa_{23}$ are set to one, change the assumptions about the baseline hazard to piecewise exponential or something more flexible, assess different effects of covariates in the transitions, and modify our proposed time-reset parameterization. In later sections, we provide a tool for assessing the sensitivity of these values, commentary on the feasibility and identifiability of estimating these models with and without these assumptions, and simulations that examine model misspecification.

### 3 CEP Quantities

Based on these proposed models for the time-to-event outcomes, we propose a method for validating a surrogate outcome using the principal stratification framework (Frangakis & Rubin, 2002). The goal of this validation procedure is to develop causal quantities that rigorously determine if a time-to-event $S$ is a valid surrogate for use in a future trial in place of $T$ by conditioning on the joint distribution of the observed and counterfactual surrogate, specifically the log cumulative hazard ratio of the time to the surrogate outcome under control versus treatment. In the setting without survival endpoints, Gilbert and Hudgens (2008) define a principal surrogate outcome for binary primary outcomes based on
the comparison of the quantities $\text{risk}_{(1)}(s_1, s_0) \equiv P(T(1) = 1|S(1) = s_1, S(0) = s_0)$ and $\text{risk}_{(0)}(s_1, s_0) \equiv P(T(0) = 1|S(1) = s_1, S(0) = s_0)$, which are functions of the counterfactual outcomes $S(z), T(z)$ under treatment $z = 1$ and $z = 0$. The condition that these must be equal for all $s_1 = s_0$ is known as average causal necessity. Average causal sufficiency is defined as $\text{risk}_{(1)}(s_1, s_0) \neq \text{risk}_{(0)}(s_1, s_0)$ for all $|s_1 - s_0| > C$ for some nonnegative constant $C$. They define the causal effect of the treatment on the true outcomes as $h(P(T(1) = 1), P(T(0) = 1))$ for some $h(\cdot)$ contrast function that satisfies $h(x, y) = 0$ if and only if $x = y$. The CEP surface is therefore equal to $h(\text{risk}_{(1)}, \text{risk}_{(0)})$ over values of $s = (s_1, s_0)$. A specific case of this is the CEP plot of $\Delta T = E(T(1) - T(0)|S(1) = S(0) = s)$ over values of $\Delta S = S(1) - S(0) = s$ when $S$ and $T$ are continuous (see Conlon, Taylor, and Elliott (2014) for further consideration of this setting). Based on these criteria, an ideal CEP plot for a valid surrogate will go through the origin and have a positive slope.

We generalize this by defining new contrasts, $\Delta T_i$ and $\Delta S_i$ for each subject in this time-to-event setting and forming a scatterplot of $(\Delta S_i, \Delta T_i)$. Each individual has a setup up to six subject-specific hazards that will be used in the CEP plot. Let $\Delta S_i = \log \frac{\Lambda^0_0(S(1), \omega_i) - \Lambda^0_0(S(0), \omega_i)}{\Lambda^1_0(S(1), \omega_i) - \Lambda^1_0(S(0), \omega_i)}$ be a measure of the treatment effect on the x-axis of the plot that depends on some time $\tau_S$ where

$$\Lambda^0_0(\tau_S | \omega_{12}) = \int_0^{\tau_S} \Lambda^0_0(t | \omega_{12}) dt$$ and
$$\Lambda^1_0(\tau_S | \omega_{12}) = \int_0^{\tau_S} \Lambda^1_0(t | \omega_{12}) dt.$$ For example, using model A,

$$\Delta S_i = \log \frac{\Lambda^0_0(S(1), \omega_i) \exp(\tau_1 \omega_i) - \Lambda^0_0(S(0), \omega_i) \exp(-\tau_1 \omega_i)}{\Lambda^1_0(S(1), \omega_i) \exp(\tau_1 \omega_i) - \Lambda^1_0(S(0), \omega_i) \exp(-\tau_1 \omega_i).}$$

For $\Delta T_i$ we will use the difference in survival probabilities evaluated at time $\tau_T$. Specifically for each individual, consider $\Delta T_i = P(T_i(1) > \tau_T | \omega_{12i}, \omega_{13i}) - P(T_i(0) > \tau_T | \omega_{12i}, \omega_{13i})$ based on the frailties in model A. We are using the notation $S(z)$ and $T(z)$ here to represent the time to the surrogate (potential) outcome and true (potential) outcome; we provide a more formal introduction of the observed data in Section 4.1.

Our $\Delta T_i$ based on OS at time $\tau_T$ can be decomposed into components based on the sum of $P(\text{do not experience the surrogate or true outcome})$ and $P(\text{experience the surrogate outcome but not the true outcome})$. More formally, this framework is similar to the likelihood for a joint illness-death model developed in Suresh et al. (2017) and for illness-death with a cure fraction proposed by Conlon, Taylor, and Sargent (2014) and Beesley and Taylor (2019). In formal notation, we are interested in the quantities

$$P(T(0) > \tau_T) = P(T(0) > \tau_T, S(0) > \tau_T) + P(T(0) > \tau_T, S(0) < \tau_T)$$

and

$$P(T(1) > \tau_T) = P(T(1) > \tau_T, S(1) > \tau_T) + P(T(1) > \tau_T, S(1) < \tau_T).$$

These quantities can be written in terms of parameters

$$\exp(-\int_0^{\tau_T} \lambda_1(u) du - \int_0^{\tau_T} \lambda_3(u) du) + \int_0^{\tau_T} \exp(-\int_0^u \lambda_2(v) dv - \int_0^u \lambda_3(v) dv) \lambda_1(u) \exp(-\int_0^{\tau_T-u} \lambda_3(v) dv) du$$

$$= \exp(-\Lambda_1(\tau_T) - \Lambda_3(\tau_T)) + \int_0^{\tau_T} \exp(-\Lambda_2(u) - \Lambda_3(u)) \lambda_1(u) \exp(-\int_0^{\tau_T-u} \lambda_3(v) dv) du.$$

After calculating $\Delta T_i$ and $\Delta S_i$ conditional on the set of $\omega_i$, we propose to create a scatterplot of $\Delta T_i$ versus $\Delta S_i$ and draw a loess or linear curve through the points to achieve a type of expected value curve. In this work, our $\gamma_0$ and $\gamma_1$ summary quantities are equal to the intercept and slope of this line (wheras these quantities may need to be redefined for a loess curve). Since the intercept and the slope of best fit line depend on $\tau_S$ and $\tau_T$ we can write these as $\Delta T_i(\tau_T) = \gamma(\tau_S, \tau_T) \Delta S_i(\tau_S).$ A good surrogate will have $\gamma(\tau_S, \tau_T) = 0$ and $\gamma(\tau_S, \tau_T) > 0$, with larger values of $\gamma(\tau_S, \tau_T)$ implying better surrogacy. $\tau_S$ and $\tau_T$ must be chosen at meaningful or sensible times. $\tau_T$ would usually be determined by the clinical context, and $\tau_S$ needs to be meaningfully less than $\tau_T$ for the surrogate to be useful. While small times for $\tau_S$ and $\tau_T$ are desirable from a trial duration standpoint, they should also be chosen such that a sufficient number of events have occurred in order to make sensible decisions about the surrogate with adequate power. Finally, for the surrogate to be relevant we would want a treatment effect on $S$, so from the CEP plot we would also assess whether the mean of $\Delta S_i$ is equal to zero.

It is also possible to use other quantities for both $\Delta S$ and $\Delta T$. Here, we have chosen this $\Delta T$ as an interpretable quantity that might be used as the true outcome in the trial that can be calculated regardless of whether $S$ has occurred. We have chosen $\Delta S$ to be directly related to the transition from baseline to the intermediate state in the illness-death model, as this is what the therapies are usually aiming to modify. Other choices for $\Delta S$ are possible where it is based on a probability
TABLE 1 Eight possible scenarios of which pathways in the illness-death models exhibit treatment effects based on the causal hazards. \( T \) denotes true and \( F \) denotes false. The right-hand column represents an intuitive notion of whether the surrogate outcome is a good surrogate for the primary outcome.

| Scenario | \( \lambda_{12}^0 = \lambda_{12}^1 \) | \( \lambda_{13}^0 = \lambda_{13}^1 \) | \( \lambda_{23}^0 = \lambda_{23}^1 \) | Surrogacy            |
|----------|-----------------|-----------------|-----------------|-----------------|
| Scenario 1 | \( T \)         | \( T \)         | \( T \)         | Null case       |
| Scenario 2 | \( F \)         | \( T \)         | \( T \)         | Perfect         |
| Scenario 3 | \( F \)         | \( T \)         | \( F \)         | Partial         |
| Scenario 4 | \( F \)         | \( F \)         | \( T \)         | Partial         |
| Scenario 5 | \( F \)         | \( F \)         | \( F \)         | Partial         |
| Scenario 6 | \( T \)         | \( F \)         | \( F \)         | Not a surrogate |
| Scenario 7 | \( T \)         | \( T \)         | \( F \)         | Not a surrogate |
| Scenario 8 | \( T \)         | \( T \)         | \( T \)         | Not a surrogate |

rather than a cumulative hazard or involves more than just the transition from baseline to the intermediate state. These will be considered in the discussion section.

3.1 Valid surrogates under an illness-death CEP curve

As our novel formulation for the CEP curve is a fairly complex function of model parameters, before considering any estimation strategies we investigate what combination of illness-death models, meaning relationship between \( S \) and \( T \), leads to CEP plots that align with an intuitive notion of whether the surrogate outcome is a good surrogate for the primary outcome (Table 1). We first consider the eight scenarios that may exist based on which transitions have treatment effects (defined as whether or not the counterfactual hazards are equal) in Table 2 and in the Supporting Information. These scenarios and the magnitude of the effects determine whether there are marginal treatment effects on the surrogate outcome and the primary outcome.

In the Supporting Information, we characterize the CEP curves under these scenarios using true generating parameter and frailty values to generate scatterplots of \( \Delta T \) and \( \Delta S \). Based on these investigations, we suggest which data scenarios should correspond to a decision that the intermediate outcome is a valid surrogate; for a perfect surrogate, the treatment effects should exist through the baseline to intermediate outcome transition only (i.e., \( \lambda_{12}^0 \neq \lambda_{12}^1 \)). In the null case with no treatment effects, Scenario 1, and Scenario 2, the estimated slope is positive, and the intercept is equal to 0. This is consistent with our consideration of the Prentice criteria, which also suggest that hazards from baseline to surrogate outcome should be nonequal (\( \lambda_{12}^0 \neq \lambda_{12}^1 \)) and the hazards from baseline to true outcome should be equal (\( \lambda_{13}^0 = \lambda_{13}^1 \)) across treatment arms.

An R shiny app is also available at https://emilyroberts.shinyapps.io/id_cep_parameters/ that allows users to characterize the CEP curve for different parameter values and under different independence or equivalence assumptions about the frailty terms. Largely, small changes in the values of \( \rho \) in the correlation matrix of the frailty terms does not have a major impact on the CEP slope and intercepts, though other settings in the online app demonstrate specific settings where these correlations may be more consequential.

We examine the marginal effects on \( S \) and \( T \) based on the average of \( \Delta S_i \) and \( \Delta T_i \) and via Kaplan–Meier curves. For Scenario 1, treatment effects on both outcomes are zero, which may correspond to a treatment not worth future investigation. For other scenarios, the marginal effect on \( T \) is somewhat small under the parameter values we are presenting. We did observe that Scenarios 3–8 (denoted as partial and nonsurrogates) produced CEP curves that did not go through the origin and therefore were invalid. We anticipated differences between perfect, partial, and nonsurrogates would be easily apparent, and while the intercepts did differ, the slope did not drastically change across the scenarios. Under the particular parameters we investigated, the slope was positive for all scenarios when the baseline hazard to \( T \) was larger after experiencing the surrogate outcome (i.e., the baseline hazard \( \lambda_{0,23}^Z > \lambda_{0,13}^Z \) so that death occurs faster after progression). In other words, the relative magnitude of the baseline hazards for gap times for a given treatment arm influences the slope and intercept of a CEP curve. Differences smaller than expected across scenarios may be explained by \( \Delta T_i \) necessarily constrained between \(-1 \) and \( 1 \) since it represents a difference in two probabilities or \( \Delta T_i \) being a complex function of model parameters (so it may not change drastically based on relatively small changes in the hazards).
### Additional scenarios

In addition to which hazards are moderated by treatment in the eight settings in Table 2, each combination can be crossed with whether \( \theta_{23} \) and \( \kappa_{23} \) are zero versus nonzero in a factorial design. We briefly considered the former and do see that incorporating nonzero values of \( \phi_{23} \) does change the CEP curve slope and intercept in the Supporting Information. While these settings taken together represent a broad range, there are many other possible scenarios that could be achieved with specific choices of the parameters. For example, even if a treatment slows the rate of progression to the surrogate outcome, it is possible that time to death after progression may be more rapid on the treatment arm. In our setting, that would be seen in a positive treatment effect on the transition from baseline to surrogate outcome, but a negative treatment effect from the surrogate to primary outcome either through increasing the baseline hazard \( \lambda_{12} \) or a positive value of \( \phi_{12} \). Another possibility exists where the treatment slows the rate of progression, corresponding to a positive treatment effect from baseline to surrogate outcome, however toxicities or side effects from the treatment effect cause death from other
causes, affecting the baseline to $T$ transition to have a negative treatment effect. More complex study designs might allow for patients to switch to the active treatment arm after experiencing the surrogate outcome, which could be potentially incorporated into our illness-death framework by reducing $\lambda_{23}^0$.

4 | LIKELIHOOD AND ESTIMATION

4.1 | Likelihood contributions

We consider a randomized clinical trial of $n$ subjects for a binary treatment $Z$. For generality, let $n_z$ denote the number of subjects in treatment arm $Z = z$ (and we may assume that $n/2$ subjects are in treatment group $Z = 1$ and $n/2$ are in treatment group $Z = 0$ since the treatment assignment is randomized and under the control of the investigator). Let $\{S_i, \delta_{Si}, T_i, \delta_{T_i}, Z_i\}$ be the observed data for subject $i$ for $i = 1, \ldots, n$. We will also consider a random or administrative censoring time $C_i$. $S_i$ denotes the time that the surrogate outcome occurs (i.e., the transition to state $S$), $T_i$ denotes the time that the primary event occurs, meaning the individual enters state $T$, and $\delta_T$ and $\delta_S$ denote the censoring indicators for $T$ and $S$ being observed. Then $\delta_{T_i} = 1$ when $T_i < C_i$ and $\delta_{Si} = 1$ when $S_i < C_i$ and $S_i < T_i$.

We can also conceptualize the data in terms of the random variables in Figure 1. Based on gap times $T_{jk}(z)$ between states, the data can also be represented as $\{T_{12i}, T_{13i}, T_{23i}, \delta_{Si}, \delta_{T_i}\}$ with $T_{23i}$ not defined when $S_i$ is not observed. In the illness-death formulation, there are four possible combinations of observable $\delta_{Si}$ and $\delta_{T_i}$. We assume that when neither event is observed, meaning $\delta_{Si} = \delta_{T_i} = 0$, then $T_{12i}(z)$ and $T_{13i}(z)$ take on the same value as being censored at $C_i$. Consider when $T$ is observed before $S$, meaning $\delta_{T_i} = 1, \delta_{Si} = 0$. Then the observed data related to $S_i$ for individual $i$ is equal to $\{T_{13i}, \delta_{Si} = 0\}$, and observed $T_i$ is based on $\{T_{12i}, \delta_{T_i} = 0\}$, while $T_{23i}$ is not defined. Now consider when only $S$ is observed, meaning $\delta_{T_i} = 0, \delta_{Si} = 1$. Then, the observed data for individual $i$ is $S_i$ based on $\{T_{12i}, \delta_{Si} = 1\}$. Assuming $T$ is not observed after the value $T_i$ takes on is censored at $\{C_i, \delta_{T_i} = 0\}$. If both $S$ and $T$ are observed with $\delta_{T_i} = 1, \delta_{Si} = 1$, then $T_{12i}$ is based on $\{T_{12i}, \delta_{Si} = 1\}$, and $T_i$ is based on $\{T_{12i} + T_{23i}, \delta_{T_i} = 1\}$. We provide the likelihood under these scenarios next.

We assume that each hazard in Figure 1 follows a Weibull distribution, so $\lambda_{jk}(u) = \exp(\gamma_{jk} u)\exp(\omega_{jk} u)$ for $\gamma_{jk} > 0$. The completedata likelihood is derived using the random variables in Figure 1 with both sets of counterfactual outcomes under the two treatment arms $T_{12}(0), T_{12}(1), T_{13}(0), T_{13}(1), T_{23}(0), T_{23}(1)$. This approach considers the joint model of the outcomes and involve all elements $\rho$ of the correlation matrix in Equation (4). Using this specification, an imputation scheme could be proposed to fill in all missing outcomes. Any relation between the potential outcomes across treatment arms for an individual in the complete data likelihood is not identified. Based on previous exploration of methods that use either the observed or the complete data likelihood (Roberts et al., 2021), using this complete data likelihood and employing imputation is not necessary to carry out the estimation procedure. Here, we will only focus on the observed data likelihood during estimation and consider each arm of the trial separately. For ease of notation, we will drop the superscript in this section as the derivations apply to both treatment arms. Any counterfactual quantities needed for calculation of the CEP curve will be described separately in Section 4. We note that $\{T_{23i}, \omega_{23i}\}$ are not defined when $\delta_{Si} = 0$ and do not contribute to the likelihood, which is the case for either the complete data or observed data likelihood.

The likelihood contributions can be written similarly to work done by Conlon, Taylor, and Sargent (2014). Conditional on the frailties and the other parameters the likelihood contribution for subject $i$ is

$$L_i = L(T_{12i}, T_{13i}, T_{23i}, \delta_{Si}, \delta_{T_i}; \omega_{12i}, \omega_{13i}, \omega_{23i}, Y_{12i}, \alpha_{12i}, \gamma_{13i}, \alpha_{13i}, \gamma_{23i}, \alpha_{23i}, \delta_{23i}, \kappa_{12i}, \kappa_{13i}, \kappa_{23i}).$$

For those who had not experienced $S$, we are in the setting where $\delta_{Si} = 0$, $T_{12i} = T_{13i}$ and $T_{23i}$ is not defined, then $L_i = \lambda_{13}(T_{13i})^\gamma_{13i} \exp(- \int_0^{T_{13i}} \lambda_{13}(u)du - \int_0^{T_{13i}} \lambda_{12i}(u)du \lambda_{23}(T_{23i}|T_{12i})^\gamma_{23i} \exp(- \int_0^{T_{23i}} \lambda_{23}(u)du)du)\lambda_{13}(T_{13i})^\gamma_{13i} \exp(- \int_0^{T_{13i}} \lambda_{13}(u)du - \int_0^{T_{13i}} \lambda_{12i}(u)du \lambda_{23}(T_{23i}|T_{12i})^\gamma_{23i} \exp(- \int_0^{T_{23i}} \lambda_{23}(u)du)du)\lambda_{13}(T_{13i})^\gamma_{13i}$

For those who experience $S$, and are either dead or alive, $\delta_{Si} = 1$, and $T_{23i}$ is defined. $\delta_{T_i}$ may be equal to either 0 or 1 depending on if the terminal event is observed:

$$L_i = \lambda_{12}(T_{12i}) \exp(- \int_0^{T_{12i}} \lambda_{12i}(u)du - \int_0^{T_{12i}} \lambda_{13i}(u)du \lambda_{23}(T_{23i}|T_{12i})^\gamma_{23i} \exp(- \int_0^{T_{23i}} \lambda_{23}(u)du)du).$$
4.2 Bayesian estimation

To facilitate estimation, we take a Bayesian approach using Markov chain Monte Carlo (MCMC). We use prior distributions similar to those suggested in Gao et al. (2012) and Sahu et al. (1997). Regression coefficients are assumed to have a diffuse normal prior (Sahu et al., 1997). We assume a Gamma($p_1, p_2$) prior for the scale parameters $\gamma_{jk}$ of the Weibull distribution, and we also assume a Gamma($p_1, p_3$) prior for the shape parameters $\alpha_{jk}$ with hyperparameters $p_1 = p_2 = p_3 = p_4 = 0.1$.

Any parameters that do not have a closed-form posterior distribution ($\alpha^2_{jk}, \gamma_{jk}, \omega^2_{jk}, \delta^2_{jk}, \kappa^2_{jk}$) are drawn using a Metropolis–Hastings step (Robert & Casella, 2004). At each iteration of the MCMC, proposed parameters are taken from a Gaussian proposal distribution $\pi$ with mean equal to the previous accepted draw. For a general parameter $\beta$ and iteration $p$ of the MCMC, we draw a proposed value of $\beta'_{p} \sim N(\beta_{p-1}, \sigma^2)$ based on using the previous iteration $\beta_{p-1}$.

The acceptance ratio is calculated as $\frac{P(\beta')}{P(\beta_{p-1})} \times \frac{\pi(\beta_{p-1})}{\pi(\beta')}$, where $P(\beta)$ represents the posterior distribution of $\beta$ and $\pi$ represents the proposal density. For a general Gaussian density, $g(\beta' | \beta_{p-1}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-1/2\sigma^2)(\beta' - \beta_{p-1})^2$ and $g(\beta_{p-1} | \beta') = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-1/2\sigma^2)(\beta_{p-1} - \beta')^2$. Based on our proposal distribution, the exponential terms in the ratio of Gaussian densities will cancel, so the proposed draw $\beta'$ is accepted with the simplified probability $\min(1, \frac{P(\beta')}{P(\beta_{p-1})})$. The variance of the proposal distribution $\sigma^2$ is tuned to obtain convergence of parameter draws and target a reasonable acceptance rate (Gelman et al., 1996). The proposal distributions have standard deviation (SD) $\sigma = 0.1$.

The frailties are also drawn using a Metropolis–Hastings step with a Gaussian proposal distribution with mean equal to the previous value and a Gaussian prior with mean zero and SD equal to 0.4. Each proposed frailty term for an individual (Gelman et al., 1996). The proposal distributions have standard deviation (SD) $\sigma = 0.1$.

Visually, we can see the hierarchy of parameters across different treatments and transitions and how the terms are related in Figure 2.

Initial estimates of the frailties may be calculated using the frailtypack or frailtyEM packages in R (Balan & Putter, 2019; R Core Team, 2021; Rondeau & Gonzalez, 2005). Parameter estimates are each drawn from the proposal distribution individually. Under the parameterization in model A, $\delta^2_{23}$ is drawn from a proposal distribution with a mean based on the estimated coefficient from a hazard model fit using observed data regressing time $T$ on time $S$, among those who experience the surrogate outcome. By doing this, $\delta^2_{13}$ and $\delta^2_{03}$ have unique starting values. The draws are accepted in blocks for the Metropolis–Hastings step. The blocks are divided into treatment arm transitions, and the parameters within a block are jointly accepted or rejected. For model A, we have blocks $\omega^2_{12}, \psi^2_{12}, \omega^2_{23}, \theta^2_{23}, \alpha^2_{12}, \alpha^2_{23}, \beta^2_{12}, \beta^2_{23}, \gamma^2_{12}, \gamma^2_{23}, \delta^2_{12}, \delta^2_{23}, \kappa^2_{12}, \kappa^2_{23}$ when all of the model parameters are being estimated.

While counterfactual draws of the frailties are not needed for the estimation procedure, they are needed to form the proposed CEP plot. As the correlations between the observed and counterfactual outcomes are nonidentified, we fix $\rho_S, \rho_T$ from the distributions in Equation (4) to draw the counterfactual frailty terms. We use correlations of 0.5 as a starting point since it is a midpoint between perfect and no correlation and then vary $\rho_S$ and $\rho_T$ for sensitivity analysis. We use the normal prior distribution and fixed $\rho_S, \rho_T$ to obtain draws of the $\omega$ estimates in the counterfactual arm from the appropriate conditional normal distributions, such as $\omega^2_{12}\omega^2_{12} \sim N(0 + \rho_S(\omega^2_{12}), 1 - \rho_S^2)$ and similarly.
$\omega_{13}^{1-z} \sim N(0 + \rho_{T}(\omega_{13}^{1-z}), 1 - \rho_{T}^2)$. We repeat the process for the other treatment arm to obtain sets of counterfactual frailties for each individual.

Based on the draws of model parameters for a given iteration of the MCMC, we estimate observed and counterfactual hazards for each individual to ultimately calculate the CEP quantities. This process is repeated for the next set of draws of model parameters and frailties for all individuals. These quantities are then averaged over MCMC iterations after a burn-in period.

## 5 SIMULATION STUDY

### 5.1 Simulation setup

Here we start with a simulation setting where frailty coefficients are equal to 1. We conduct a simulation with 200 replicated data sets and $n = 600$. Data are generated under simple settings that follow the $\theta$ parameterization shown in model A. The true values of the parameters are shown in the simulation results in the first row of the table of results. Survival times are simulated based on a Weibull baseline hazards specification (Austin, 2012). We generate treatment effects by differing the scale parameters between arms, meaning $\gamma_{1j} \neq \gamma_{0j}$. We simulate the frailties to have mean 0 and an SD of 0.4 and assume that $\omega_{13}^{1-z} = \omega_{23}^{1-z}$ in our primary results settings.

We conduct the estimation procedure described in Section 3 from our eight simulation scenarios, highlighting Scenario 1 with no marginal treatment effects (a null setting), Scenario 2 where there is a treatment effect only on the surrogate outcome (labeled as a perfect surrogate), and Scenarios 3–8 where treatment effects exist such that we do not expect the intermediate outcome to be a surrogate. Because of nonidentifiability due to the close link between the baseline hazard, frailties, and coefficients associated with the frailties, we assume during estimation that all $\kappa_{j, k}^z = 1$. In the Supporting Information, we also conduct sensitivity analyses by varying the assumptions that $\omega_{12}^{z} \perp \omega_{13}^{z}$ and $\omega_{13}^{z} = \omega_{23}^{z}$. There we assume that either $\omega_{12}^{z} \perp \omega_{13}^{z} \perp \omega_{23}^{z}$ or that all three frailties are correlated within a given counterfactual treatment arm. In this case, we assume $\rho_{T1} = \rho_{T4} = 0.95$ and $\rho_{T3} = \rho_{T2} = \rho_{ST} = \rho_{S} = 0.5$ and set $\tau_{S} = 1$ and $\tau_{T} = 5$.f
5.2 Simulation results

In this section, we show results of the estimated model parameters as well as validation quantities, the intercept $\gamma_0$, and slope $\gamma_1$. The estimation of the $\gamma_0$ and $\gamma_1$ quantities are calculated from fitting a linear best fit line through the CEP cloud at each iteration and reporting the posterior mean of these quantities for each simulated data set. Parameter estimates are based on the posterior means and corresponding measures of variability; the average estimated standard error (SE) and the SD of the posterior means are shown for the model parameters. We run the simulations for 3000 iterations with 900 burn in draws. In addition to trace plots of the parameter draws, we assess the empirical mean and SD of the estimated frailty terms over the iterations.

In Figure 3, we show the CEP curve conditional on estimated frailties for one data set under Scenario 2. Each point is the posterior mean of $(\Delta S_i, \Delta T_i)$ across MCMC iterations. The posterior values of the slope and intercept are shown, which convey the amount of variability based on the posterior coordinates of $(\Delta S_i, \Delta T_i)$ for each individual $i$. We see that the estimated slope and intercept correctly meet our criteria of a valid surrogate under our proposed set of model assumptions. Though there is substantial variability in the estimates of $\gamma_0$ and $\gamma_1$, the respective posterior mean and credible intervals are $-0.018 (-0.078, 0.042)$ and $0.049 (0.020, 0.078)$ for this data set. Furthermore, there is a marginal effect of the treatment on both $S$ and $T$ for this data set, as denoted by the nonzero position of the dashed lines.

In the main set of simulations in Table 2, the identified parameters are estimated fairly well and seem to converge based on the assumptions we have made. We observe that the distribution of the estimated frailty terms can deviate from the generating distribution with mean zero and fixed variance. While our method involves prior and proposal distributions for the frailties, we are not directly enforcing any assumptions about the mean or variability of the frailty parameters during the estimation algorithm. The shape of the likelihood for frailty terms, particularly $\omega_{12}$ terms for individuals with $\delta_{Si} = 0$, seems to be fairly flat, so the draws move around considerably during the algorithm. In these considered simulations, the credible intervals around $\gamma_1, \gamma_0$ are somewhat wide for all scenarios. Since an ideal surrogate will have values $\gamma_0 = 0$ and $\gamma_1 > 0$, too much uncertainty can make it difficult to determine the value of the surrogate.

We conducted a study of model misspecification via simulation by generating the survival data under a Gompertz baseline hazard structure. With these data, we assume all baseline hazards follow a Weibull distribution during our estimation procedure. Results found in the Supporting Information showed differences in estimated $\gamma_{jk}$ and $\sigma_{jk}$ parameters as expected. Importantly, the general conclusions about surrogacy (i.e., the estimates and credible intervals of $\gamma_0$ and $\gamma_1$ covering zero) were similar, however. In our sensitivity analyses about the assumptions on the frailty terms, shown in the
Supporting Information, we see some sensitivity, such as increased variability in the subject-specific points. How these factors influence the CEP curves should be investigated under trial-specific contexts.

6 | PROSTATE CANCER EXAMPLE

Our motivating clinical study is a phase III, randomized trial for men with prostate cancer, NRG Oncology, Radiation Therapy Oncology Group (RTOG) trial 9601 (Shipley et al., 2017). The trial features 760 men with recurrently or persistently elevated prostate-specific antigen (PSA) levels whose prostate was initially removed by prostatectomy. The two treatments being compared are postprostatectomy radiation therapy with or without antiandrogen therapy. There are 384 and 376 men in each treatment arm. The two survival outcomes of interest are time to distant metastasis, defined as radiographic evidence of metastatic cancer, and OS. Notably, composite outcomes such as metastasis-free survival (MFS) are often evaluated. It has been previously established by The Intermediate Clinical Endpoints in Cancer of the Prostate (ICECaP) that MFS is a valid surrogate for OS in the setting of the initial treatment for localized prostate cancer (Xie et al., 2017). Others have evaluated if MFS is a valid surrogate when assessing the impact of antiandrogen therapy in recurrent prostate cancer following postprostatectomy salvage radiation therapy (Jackson et al., 2022). However, within our illness-death framework we consider time to distant metastasis and time to death separately for our primary analysis.

We show in Figure 4 the Kaplan–Meier curves for the intermediate and true outcomes without considering the semi-competing risk as well as the curve for the transition from the surrogate to primary outcome for those who experienced distant metastasis. S may be censored because it was not observed during the study period or because the terminal event occurred first. In the Supplemental Information, in Figure S6, we also present the cumulative incidence curve for S considering death as a semicompeting risk based on the nonparametric Aalen–Johansen estimate of the cumulative incidence function from the mstate package (Putter, 2011). The plots show that the addition of antiandrogen therapy decreases the hazard of distant metastases and increases the survival probability, but after metastases the survival probability is reduced and does not appear to be greatly influenced by whether the antiandrogen therapy was part of the treatment.

6.1 | Existing methods

We calculate the number of individuals who go through each transition and experience the events in our illness-death models. In total, 156 patients experienced distant metastases, and 239 total deaths were observed between the two arms. These numbers are shown in Figure 5. We consider the z = 1 group to be the treatment group for salvage radiation therapy with antiandrogen therapy, and the z = 0 represents the group treated without antiandrogen therapy. There is a significant treatment effect of the additional antiandrogen therapy on time to distant metastasis using a parametric hazard model with a Weibull baseline hazard (HR = 0.622, p = 0.004) and a marginally significant treatment effect on OS when considering the cause-specific hazard (HR = 0.722, p = 0.049).

As noted in the introduction, other methods that can assess surrogacy in this data setting rely on different frameworks and assumptions. To consider the Prentice criteria, we fit a model for OS adjusting for the occurrence of distant metastases as a time-dependent covariate. We found that the effect was attenuated toward null (HR = 0.890, p = 0.592) and no longer statistically significant. We explored a mediation approach (Weir et al., 2021) and estimated the surrogate outcome provided 44% proportion mediated of the treatment effect at the evaluation time of 5 years. While this metric relies on different modeling and causal assumptions than ours, including its use of a semiparametric Cox model and using only one time point for evaluation (S time to distant metastases is included as a covariate in the model for the terminal event), these methods suggest that the distant metastases may have value as a surrogate outcome.

6.2 | Proposed surrogacy evaluation

As localized prostate cancer is a relatively slowly progressive disease compared to most other cancers, it is quite common in this setting for trials to have long follow-up since cancer-related events (such as distant metastases) are not particularly common outcomes. Based on the Kaplan–Meier curves and typical survival times, we chose τS = 5 years and τT = 8 years, though these could be varied in sensitivity analyses, or a grid search or optimization could potentially be done to choose these values. We show an estimated CEP curve based on model A using T12 as a time-varying covariate where we assume
**FIGURE 4** Kaplan–Meier curves for the intermediate and true outcome demonstrating significant treatment effects for the prostate cancer trial. We also show the Kaplan–Meier curve for the transition from the surrogate to primary outcome among those who experienced the surrogate outcome.

**FIGURE 5** Counterfactual illness-death models for baseline, illness ($\tilde{S}$), and death ($\tilde{T}$) with the number of individuals experiencing the events in each transition for the prostate cancer trial.

\[ \kappa^2_{12} = \kappa^2_{13} = \kappa^2_{23} = 1. \] Table 3 shows the posterior mean and corresponding 95% credible interval for each parameter being estimated. We plot the posterior mean of $\Delta S_i$ and $\Delta T_i$ for each individual across iterations in a CEP plot. We also show the estimated slope and intercept lines on the CEP curve for each iteration of the MCMC chain to assess the variability of the estimates of these validation quantities.

Based on this example data set and CEP curve in Figure 6, the vertical and horizontal lines for the marginal treatment effects are separated from zero. Across individuals, the average $\Delta S_i$, our measure of the treatment effect, is 0.18, and the
TABLE 3  Parameter estimates for the prostate cancer data example. The posterior mean and estimated standard error (SE) are shown for each parameter. All $x_{jk}$ are set to 1.

| Parameter | $\gamma_0$ | $\gamma_1$ | $\gamma_{012}$ | $\gamma_{013}$ | $\gamma_{023}$ | $\gamma_{112}$ | $\gamma_{113}$ | $\gamma_{123}$ |
|-----------|-------------|-------------|----------------|----------------|----------------|----------------|----------------|----------------|
| Posterior mean | 0.025 | 0.047 | 0.024 | 0.019 | 0.132 | 0.017 | 0.018 | 0.116 |
| SE | 0.035 | 0.012 | 0.003 | 0.002 | 0.040 | 0.002 | 0.002 | 0.049 |

Parameter

| Parameter | $\sigma_0^2$ | $\sigma_1^2$ | $\sigma_{012}^2$ | $\sigma_{013}^2$ | $\sigma_{023}^2$ | $\sigma_{112}^2$ | $\sigma_{113}^2$ | $\sigma_{123}^2$ |
|-----------|-------------|-------------|----------------|----------------|----------------|----------------|----------------|----------------|
| Posterior mean | 0.067 | 0.045 | 0.959 | 0.940 | 1.039 | 0.923 | 0.945 | 1.309 |
| SE | 0.033 | 0.033 | 0.044 | 0.041 | 0.105 | 0.046 | 0.042 | 0.258 |

![Illness-Death CEP Curve for Prostate Cancer Trial](image)

FIGURE 6  Causal effect predictiveness plot for the motivating prostate cancer trial data set. Each point represents the posterior mean of $\Delta S_i$ and $\Delta T_i$ for an individual. The collection of linear best fit lines in gray represent the posterior slope $\gamma_1$ and intercept $\gamma_0$ evaluated at each iteration of the Markov chain Monte Carlo (MCMC). The posterior marginal effects on $S$ and $T$ are shown in the dotted lines. CEP, causal effect predictiveness.

average $\Delta T_i$ is 0.04. The posterior mean for the intercept term $\gamma_0$ is 0.025 with 95% credible interval ($-0.044$, $0.094$). For the slope $\gamma_1$, the posterior mean is 0.047 with 95% credible interval (0.002, 0.092). Based on these estimates, we would conclude that the slope $\gamma_1$ is positive, and the estimated intercept $\gamma_0$ is near zero since the credible interval for $\gamma_0$ does include 0. These results would indicate that the surrogate seems valid, though the credible interval for $\gamma_0$ is somewhat wide. We also conducted a sensitivity analysis where instead of assuming $\omega_z^{13} = \omega_z^{23}$ and that $\omega_z^{12} \perp \omega_z^{13}$, we assumed that all six counterfactual frailties were correlated within an individual. These results gave generally similar conclusions, though the validation procedure with fewer equality assumptions had larger variability as expected.

7  | DISCUSSION AND CONCLUSION

In this work, we have considered how to validate surrogate markers when trial outcomes are time-to-event using principal stratification and illness-death models. We believe the illness-death framework is foundational to modeling these data, though a single, optimal estimand corresponding to the model is less obvious. We have provided examples and an online app to explore CEP curves under different data settings. While the values of the CEP curve can be written in a closed, analytic form when the outcomes are Gaussian in previous work (Conlon, Taylor, & Elliott, 2014; Roberts et al., 2021), it is necessary to define and empirically assess what an ideal CEP curve looks like for time-to-event data. A novel distinction in this work is that in the Gaussian case, the CEP conditions on $S_i(1) - S_i(0) = s$, where the conditioning is on a contrast between potentially observable values, $S_i(1)$ and $S_i(0)$. In this paper, we are looking at the contrast between $\Lambda_{12i}^{13} = \Lambda_{12i}^{13}$, which is a contrast between distributions.

While not the case in our considered scenarios, extrapolation may be required to determine if the CEP curve goes through the origin of the plot depending on the size of the treatment effect on the surrogate outcome. The subject-specific points may not appear in all four quadrants of the plot. There is an interesting connection regarding individual-specific $\Delta S_i$ and $\Delta T_i$ within the quadrants of the graph that has been considered across trials in the meta-analytic setting (Elliott et al., 2015). In particular, certain subject-specific coordinates may suggest that the treatment has a beneficial effect on the
surrogate outcome but a detrimental effect on the true outcome for certain individuals. This may be informative when considering the possibility of the surrogate paradox (VanderWeele, 2013).

There are several areas for sensitivity analyses and exploration of identifiability for surrogacy validation (Ghosh, 2012). While the variance of the frailty should be identifiable by including sufficient covariates (Gao, 2012; Putter & van Houwelingen, 2015), it may still be difficult to accurately estimate frailty terms in a complex model. In our proposed models, we include a prior distribution for the variance of the frailty terms but do not assume the variance is known. Since allowing for too much flexibility in the models may result in nonidentifiability of parameters, this can lead to computational problems when trying to estimate the coefficients associated with the frailties. We believe our assumptions that \( \kappa_{jk} = 1 \) or that \( \omega_{13} = \omega_{23} \) about the frailty terms are justifiable for this data example. They also help with computation during estimation, but they are still potentially strong assumptions. Relaxing the assumption that the frailties going into the \( \tilde{T} \) state are equal (i.e., \( \omega_{13} = \omega_{23} \)) may impact identifiability since there will be less information available to estimate these terms. To the extent that frailties can be estimated for one event time per person, the data might inform these assumptions (e.g., the assumption is testable to the extent that frailties can be estimated well). We might try to assess the identifiability of frailty terms in the proposed causal model by comparing the prior and posterior distributions for the frailty terms (Gao, 2012). Other convergence metrics can be used to assess the convergence of the parameters, and more complex algorithms or different distributional assumptions about the frailties may alleviate computational problems (Clayton, 1991; Wen et al., 2016, for example) for example. For assessment of robustness, our models can be evaluated under further model misspecification. To increase the flexibility of the method, we could also consider fitting a nonlinear loess curve through the points on the CEP plot as opposed to a linear fit. We can compare our proposed methods to copula models (Taylor et al., 2015). These particular Gaussian copula models have potential of extending the closed-form correlation structure.

Some covariates in the dataset are also available, including PSA values at the time of randomization, Gleason score, and age in grouped categories. In previous work (Roberts et al., 2021), the validation results with covariates would be specific to a certain subgroup of patients; a postestimation step was needed to marginalize over the covariates. If we fit conditional surrogacy validation models in the proposed setting, the CEP scatterplot should still have a marginal interpretation since each point in the scatterplot represents a different person, and each has their own set of covariate values. The potential advantages of using covariates, either marginally or as interactions, can be explored more thoroughly in future work. It is likely that controlling for covariates will change the estimated frailties, as frailty terms capture unexplained heterogeneity in treatment effects which would then be partially explained by the covariates. Based on these analyses, we could also determine if the surrogate is valid for certain subgroups of people (Roberts et al., 2021). Different covariates may be more important in different transition models. For example, we may expect age to be more important for the direct transition from baseline to death, while baseline PSA and Gleason score will likely be more important for time to distant metastases. Model selection could lower the number of parameters to estimate (Reeder et al., 2022).

In the future, we can consider changing our model parameterization from our proposal to use a time-varying covariate in the transition model from the surrogate outcome to the primary outcome to the alternative model B or a different structure. We may extend beyond the proposed illness-death model to a different or more complex multistate model depending on the outcomes being evaluated. In different disease areas, consideration about individuals being cured may be appropriate (Conlon, Taylor, & Sargent, 2014). We have assumed here that time to the surrogate outcome is known, but it may be subject to interval censoring (Zeng et al.). In some cases, we may even have exact information about time to \( T \) based on death registries without knowing if the surrogate outcome occurred (Beesley & Taylor, 2019). Different models, definitions of the outcome, and corresponding \( \Delta S \) may change our determination whether the surrogate is valid, and the assumptions made about the models and frailties may be more appropriate for certain contexts.

While we believe the illness-death model is natural for modeling these data, different estimands could be considered for validation. In the CEP plot, we have used the ratio of the cumulative hazards on the horizontal axis as a measure of the treatment effect on the surrogate outcome. This was chosen because it is explicitly related to the transition from the baseline state to the state of experiencing the surrogate, and in most settings, including our prostate cancer one, the primary way in which the treatment is expected to work is by preventing or delaying the occurrence of the events in the intermediate state. There are other possible choices for what to use for \( \Delta S \) on the horizontal axis. One would be based on the difference in the cumulative incidence of the surrogate outcome by time \( \tau_S \) between the two arms, another could be based on the composite outcome of either the surrogate outcome or the primary outcome occurring by time \( \tau_S \). Both of these can be calculated from the illness-death model parameter estimates, but both are also impacted by the transition rate from the baseline state to the terminal state.
It is interesting to consider this illness-death framework when $\Delta S_i$ is based on a composite endpoint of the surrogate and primary outcome. For example, in the prostate cancer setting, distant metastases-free survival has been considered as a surrogate outcome for OS. We briefly considered this possibility in the Supporting Information. Other potential surrogates have been considered such as biochemical recurrence or time to local recurrence, and an alternative true clinical outcome could be prostate cancer-specific survival. In our setting, it is likely that individuals may only die from prostate cancer if they experience distant metastases, so there may be fewer individuals transitioning directly from baseline to cancer-specific death compared to baseline to death from other causes.

In this paper, we have considered the situation of a single trial, in contrast to the meta-analysis setting in which data from multiple trials are analyzed. We developed an approach to assess whether the surrogate outcome is valid for replacing the primary outcome from a causal perspective. The hope would be that if the intermediate outcome is a good surrogate for the primary outcome in one trial, then it would also be a good surrogate for the surrogate outcome in other trials with similar treatments. The fact that the surrogacy measure is based on causal concepts, not just measures of association, may make it more likely to transport from one trial to the next. In previous work, in a different data setting, we find similar CEP plots across four different treatments comparisons (Taylor et al., 2015). Furthermore, in this paper, a mechanistic approach to disease progression, implicit in the illness-death model, has been taken. This illness-death structure does transport from one trial to the next, so may believe that our approach will assess surrogates in a way that is more generalizable across treatments than methods that rely on composite endpoints, that do not require the illness-death structure. This comparison, concept of transportability, and potential need for replication across several trials remain as future work (Pearl & Bareinboim, 2022).

There are other directions for extending this work, particularly when considering the overlap of causal inference and survival analysis and delicate interpretation of hazard ratios and multiple time-to-event outcomes. Others (Gran et al., 2015; Valeri et al., 2023) explore causal tools for multistate models such as inverse probability weighting, G-computation, and manipulating hypothetical transition intensities. Still more directions for future work are to formally compare the proposed models with the similar structures of the Prentice criteria, mediation strategies, or other causal methods.

ACKNOWLEDGMENTS
We would like to acknowledge NRG for the RTOG data and thoughtful feedback by Drs. Matthew Schipper, Walter Dempsey, and Ben Hansen on this work. We also acknowledge National Institutes of Health (NIH) Grant CA129102.

CODE AND SOFTWARE
Code for simulation studies is available at https://github.com/emilykroberts/Illness-Death-Surrogacy.

CONFLICT OF INTEREST STATEMENT
The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT
Research data are not shared. A pseudo data set is provided to demonstrate the data structure.

OPEN RESEARCH BADGES
This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge “Reproducible Research” for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to data confidentiality issues.

ORCID
Emily K. Roberts https://orcid.org/0000-0002-5838-9691

REFERENCES
Austin, P. C. (2012). Generating survival times to simulate Cox proportional hazards models with time-varying covariates. Statistics in Medicine, 31(29), 3946–3958.
Balan, T. A., & Putter, H. (2019). frailtyEM: An R package for estimating semiparametric shared frailty models. *Journal of Statistical Software, 90*(7), 1–29.

Beesley, L. J., & Taylor, J. M. (2019). EM algorithms for fitting multistate cure models. *Biostatistics, 20*(3), 416–432.

Bühler, A., Cook, R. J., & Lawless, J. F. (2022). Multistate models as a framework for estimand specification in clinical trials of complex processes. *arXiv preprint arXiv:2209.13658. https://arxiv.org/abs/2209.13658*

Clayton, D. G. (1991). A Monte Carlo method for Bayesian inference in frailty models. *Biometrics, 47*(2), 467–485.

Comment, L., Mealli, F., Haneuse, S., & Zigler, C. (2019). Survivor average causal effects for continuous time: A principal stratification approach to causal inference with semicompeting risks. *arXiv preprint arXiv:1902.09304.*

Conlon, A. S. C., Taylor, J. M. G., & Elliott, M. R. (2014). Surrogacy assessment using principal stratification when surrogate and outcome measures are multivariate normal. *Biostatistics, 15*(2), 266–283.

Conlon, A. S. C., Taylor, J. M. G., & Sargent, D. J. (2014). Multi-state models for colon cancer recurrence and death with a cured fraction. *Statistics in Medicine, 33*(10), 1750–1766.

Elliott, M. R., Conlon, A. S. C., Li, Y., Kaciroti, N., & Taylor, J. M. G. (2015). Surrogacy marker paradox measures in meta-analytic settings. *Biostatistics, 16*(2), 400–412.

Emura, T., Nakatochi, M., Murotani, K., & Rondeau, V. (2017). A joint frailty-copula model between tumour progression and death for meta-analysis. *Statistical Methods in Medical Research, 26*, 2649–2666.

Frangakis, C., & Rubin, D. (2002). Principal stratification in causal inference. *Biometrics, 58*(1), 21–29.

Gabriel, E. E., & Gilbert, P. B. (2014). Evaluating principal surrogate endpoints with time-to-event data accounting for time-varying treatment efficacy. *Biostatistics, 15*(2), 251–265.

Gabriel, E. E., Sachs, M. C., & Gilbert, P. B. (2015). Comparing and combining biomarkers as principal surrogates for time-to-event clinical endpoints. *Statistics in Medicine, 34*(3), 76–105. [http://www.R-project.org/]

Huang, Y.-T. (2021). Causal mediation of frailty in multistate cure models. *Biostatistics, 20*(3), 416–432.

Gelman, A., Roberts, G. O., & Gilks, W. R. (1996). Efficient metropolis jumping rules. *Bayesian Statistics, 5*, 599–608.

Ghosh, D. (2012). A causal framework for surrogate endpoints with semi-competing risks data. *Statistics & Probability Letters, 82*(11), 1898–1902.

Gilbert, P., & Hudgens, M. (2008). Evaluating candidate principal surrogate endpoints. *Biometrics, 64*(4), 1146–1154.

Gran, J. M., Lie, S. A., Oyeflaten, I., Borgan, Ø., & Aalen, O. (2015). Causal inference in multi-state models sickness absence and work for 1145 participants after work rehabilitation. *BMC Public Health, 15*(1), 1–16.

Hsu, J. Y., Kennedy, E. H., Roy, J. A., Stephens-Shields, A. J., Small, D. S., & Joffe, M. M. (2015). Surrogate markers for time-varying treatments and outcomes. *Clinical Trials, 12*(4), 309–316.

Huang, Y.-T. (2021). Causal mediation of semicompeting risks. *Biometrics, 77*(4), 1143–1154.

Jackson, W. C., Tang, M., Schipper, M., Sandler, H. M., Zumsteg, Z. S., Elstathiu, J. A., Shipley, W. U., Seiferheld, W., Lukka, H., Bahary, J. P., Zietman, A. L., Pisansky, T. M., Zeitzer, K. L., Hall, W. A., Dess, R. T., Lovett, R. D., Balogh, A., Feng, F. Y., & Spratt, D. E. (2022). Biochemical failure is not a surrogate end point for overall survival in recurrent prostate cancer: Analysis of NRG oncology/RTOG 9601. *Journal of Clinical Oncology, 40*(27), 3172–3179.

Joffe, M. M., & Greene, T. (2009). Related causal frameworks for surrogate outcomes. *Biometrics, 65*(2), 530–538.

Kemp, R., & Prasad, V. (2017). Surrogate endpoints in oncology: When are they acceptable for regulatory and clinical decisions, and are they currently overused? *BMC Medicine, 15*(1), 1–7.

O’Quigley, J., & Flandre, P. (2012). Discussion on “Meta-analysis for surrogacy: Accelerated failure time models and semicompeting risks modeling”. *Biometrics, 68*(1), 242–245.

Parast, L., Cai, T., & Tian, L. (2017). Evaluating surrogate marker information using censored data. *Statistics in Medicine, 36*(11), 1767–1782.

Pearl, J., & Bareinboim, E. (2022). External validity: From do-calculus to transportability across populations. *Statistical Science, 29*(4), 579–595.

Prentice, R. L. (1989). Surrogate endpoints in clinical trials: Definition and operational criteria. *Statistics in Medicine, 8*(4), 431–440.

Puter, H. (2011). *Tutorial in biostatistics: Competing risks and multi-state models. Analyses using the mstate package.* Leiden University Medical Center, Department of Medical Statistics and Bioinformatics [Online Tutorial].

R Core Team. (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. [https://www.R-project.org/]

Roberts, E. K., Elliott, M. R., & Taylor, J. M. (2023). Solutions for surrogacy validation with longitudinal outcomes for a gene therapy. *Biometrics, 79*(3), 1840–1852.
Rondeau, V., & Gonzalez, J. R. (2005). Frailtypack: A computer program for the analysis of correlated failure time data using penalized likelihood estimation. *Computer Methods and Programs in Biomedicine, 80*(2), 154–164.

Sahu, S. K., Dey, D. K., Aslanidou, H., & Sinha, D. (1997). A Weibull regression model with gamma frailties for multivariate survival data. *Lifetime Data Analysis, 3*(2), 123–137.

Shipley, W. U., Seiferheld, W., Lukka, H. R., Major, P. P., Heney, N. M., Grignon, D. J., Sartor, O., Patel, M. P., Bahary, J.-P., Zietman, A. L., Pisansky, T. M., Zeitzer, K. L., Lawton, C. A. F., Feng, F. Y., Lovett, R. D., Balogh, A. G., Souhami, L., Rosenthal, S. A., Kerlin, K. J., Dignam, J. J., … for the NRG Oncology RTOG. (2017). Radiation with or without antiandrogen therapy in recurrent prostate cancer. *New England Journal of Medicine, 376*(5), 417–428.

Sofeu, C. L., Emura, T., & Rondeau, V. (2020). A joint frailty-copula model for meta-analytic validation of failure time surrogate endpoints in clinical trials. *Biometrical Journal, 63*(2), 423–446.

Suresh, K., Taylor, J. M., Spratt, D. E., Daignault, S., & Tsodikov, A. (2017). Comparison of joint modeling and landmarking for dynamic prediction under an illness-death model. *Biometrical Journal, 59*(6), 1277–1300.

Tanaka, S., Matsuyama, Y., & Ohashi, Y. (2017). Validation of surrogate endpoints in cancer clinical trials via principal stratification with an application to a prostate cancer trial. *Statistics in Medicine, 36*(19), 2963–2977.

Taylor, J. M. G., Conlon, A. S. C., & Elliott, M. R. (2015). Surrogacy assessment using principal stratification with multivariate normal and Gaussian copula models. *Clinical Trials, 12*(4), 317–322.

Valeri, L., Proust-Lima, C., Fan, W., Chen, J. T., & Jacqmin-Gadda, H. (2023). A multistate approach for the study of interventions on an intermediate time-to-event in health disparities research. *Statistical Methods in Medical Research, 32*(8), 1445–1460.

Vandenberghhe, S., Duchateau, L., Slaets, L., Bogaerts, J., & Vansteelandt, S. (2018). Surrogate marker analysis in cancer clinical trials through time-to-event mediation techniques. *Statistical Methods in Medical Research, 27*(11), 3367–3385.

VanderWeele, T. J. (2013). Surrogate measures and consistent surrogates. *Biometrics, 69*(3), 561–565.

Weir, I. R., Rider, J. R., & Trinquart, L. (2021). Counterfactual mediation analysis in the multistate model framework for surrogate and clinical time-to-event outcomes in randomized controlled trials. *Pharmaceutical Statistics, 21*(1), 163–175.

Xie, W., Regan, M. M., Buyse, M., Halabi, S., Kantoff, P. W., Sartor, O., Soule, H., Clarke, N. W., Collette, L., Dignam, J. J., Pizzi, K., Paruleker, W. R., Sandler, H. M., Sydes, M. R., Tombal, B., Williams, S. G., Sweeney, C. J., & ICECaP Working Group. (2017). Metastasis-free survival is a strong surrogate of overall survival in localized prostate cancer. *Journal of Clinical Oncology, 35*(27), 3097–3104.

Xu, J., Kalbfleisch, J. D., & Tai, B. (2010). Statistical analysis of illness-death processes and semicompeting risks data. *Biometrics, 66*(3), 716–725.

Xu, Y., Scharfstein, D., Moeller, P., & Daniels, M. (2020). A Bayesian nonparametric approach for evaluating the causal effect of treatment in randomized trials with semi-competing risks. *Biostatistics, 21*(1), 34–49.

Zhang, J. L., & Rubin, D. B. (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by “death”. *Journal of Educational and Behavioral Statistics, 28*(4), 353–368.

Zhang, Y., Chen, M. H., Ibrahim, J. G., Zeng, D., Chen, Q., Pan, Z., & Xue, X. (2014). Bayesian gamma frailty models for survival data with semi-competing risks and treatment switching. *Lifetime Data Analysis, 20*(1), 76–105.

**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:** Roberts, E. K., Elliott, M. R., & Taylor, J. M. G. (2024). Surrogacy validation for time-to-event outcomes with illness-death frailty models. *Biometrical Journal, 66*, 2200324. https://doi.org/10.1002/bimj.202200324