Biostatistics (2022) **00**, 00, *pp.* 1–19
https://doi.org/10.1093/biostatistics/kxac044

Time-to-event surrogate endpoint validation using mediation analysis and meta-analytic data

QUENTIN LE COËNT*

Department of Biostatistics, Bordeaux Population Health Research Center, INSERM U1219, 146 rue Léo Saignat, 33076 Bordeaux, France

quentin.le-coent@u-bordeaux.fr

CATHERINE LEGRAND

ISBA/LIDAM, UCLouvain, 20 Voie du Roman Pays, B-1348 Louvain-la-Neuve, Belgium

VIRGINIE RONDEAU

Department of Biostatistics, Bordeaux Population Health Research Center, INSERM U1219, 146 rue Léo Saignat, 33076 Bordeaux, France

Summary

With the ongoing development of treatments and the resulting increase in survival in oncology, clinical trials based on endpoints such as overall survival may require long follow-up periods to observe sufficient events and ensure adequate statistical power. This increase in follow-up time may compromise the feasibility of the study. The use of surrogate endpoints instead of final endpoints may be attractive for these studies. However, before a surrogate can be used in a clinical trial, it must be statistically validated. In this article, we propose an approach to validate surrogates when both the surrogate and final endpoints are censored event times. This approach is developed for meta-analytic data and uses a mediation analysis to decompose the total effect of the treatment on the final endpoint as a direct effect and an indirect effect through the surrogate. The meta-analytic nature of the data is accounted for in a joint model with random effects at the trial level. The proportion of the indirect effect over the total effect of the treatment on the final endpoint can be computed from the parameters of the model and used as a measure of surrogacy. We applied this method to investigate time-to-relapse as a surrogate endpoint for overall survival in resectable gastric cancer.

Keywords: Joint modeling; Mediation analysis; Meta-analysis; Surrogacy; Time-to-event.

1. Introduction

Given the continued development of new and more effective therapies in oncology, the use of standard endpoints in clinical trials may affect the feasibility of these trials. For example, a significant increase in overall survival requires long-term follow-up of patients to ensure adequate statistical power and increases

*To whom correspondence should be addressed.

the duration and cost of the trial. The use of surrogate endpoints is therefore an important approach and a possible way to enable such otherwise unfeasible studies.

These surrogates or intermediate endpoints are faster to measure, potentially more frequent, and less costly to obtain. They would shorten the duration of the trial and allow new treatments to be available to patients more quickly. However, to maintain the validity of the clinical trial with respect to the final endpoint, statistical conclusions about the efficacy of the treatment on the surrogate endpoint should be extendable to the final endpoint. Consequently, putative surrogate endpoints should be statistically validated as good surrogates for the final endpoint before being used in an upcoming trial. For example, time-to-progression, defined as the duration between randomization and disease progression, is a commonly used endpoint in phase II/III oncology clinical trials. However, this endpoint has been shown to be a poor surrogate for overall survival in some cases (Burzykowski *and others*, 2008), while it has been shown to be promising in others (Lee *and others*, 2016). Therefore, it is of interest to use surrogate endpoints that have been properly validated as such.

Several approaches have been proposed to investigate the quality of an endpoint as a surrogate. First, Prentice (1989) defined a surrogate as a variable for which the effect of the treatment on the final endpoint is null when conditioned upon. Freedman *and others* (1992) proposed a measure of the proportion of treatment explained by the surrogate. Frangakis and Rubin (2002) used the causal inference framework to study a potential surrogate using principal stratification. An approach using meta-analytic data was proposed by Daniels and Hughes (1997) and Buyse *and others* (2000). Joffe and Greene (2009) provided an overview of existing methods and divided them into two categories: methods based on "effects" or on "associations."

Currently, methods based on data from meta-analyses are widely used and are the most common methods for analyzing potential surrogate endpoints (Buyse *and others*, 2016). The purpose of these methods is to study the association between a surrogate endpoint and a final endpoint at two different levels. Firstly, the association between the surrogate endpoint and the final endpoint is statistically quantified at the individual level. Secondly, the association between the treatment effect on the surrogate endpoint and the treatment effect on the final endpoint can be determined at the study level using meta-analytic data from multiple trials. A surrogate endpoint is considered to be of interest if the association is strong at both levels. The definition of these associations depends on the type of the endpoints, and these meta-analytic approaches have been developed for different types of surrogates and endpoints. An overview of the different proposed methods can be found in Buyse and Molenberghs (2005). However, because a surrogate is always a post-treatment variable, it may be confounded by post-treatment factors. In such cases, potential associations between surrogates and the final outcomes that are highlighted may not have a causal interpretation.

Causal inference methods, including mediation analysis, have therefore been proposed for validating surrogate endpoints (Zigler and Belin, 2012; Conlon *and others*, 2014; Vandenberghe *and others*, 2018). In this article, we focus on mediation analysis whose purpose is to decompose the total effect of the treatment on a (final) endpoint into a direct effect and an indirect effect through a mediator (here, the surrogate). The indirect effect of the treatment on the final endpoint through the surrogate can itself be decomposed into two parts: a direct effect of the treatment on the mediator and a direct effect of the mediator on the final endpoint. Recently, Zheng and Liu (2021) have proposed an approach to study the amount of treatment effect on a time-to-event outcome going through a longitudinal biomarker using a joint modeling approach. Our purpose is to consider the situation in which both the final endpoint and the surrogate are censored time-to-event and to extend the mediation analysis approach to account for meta-analytic data. A meta-analytic setting introduces heterogeneity into the data and strengthens the validation process of a surrogate compared with validation based on data from a single study.

This article is organized as follows. In Section 2, we introduce the causal and mediation analysis framework used to define our measure of surrogacy. We also define a joint model for time-to-event that

accounts for the meta-analytic nature of the data through trial-level random effects. In Section 3, we present a simulation study to investigate the finite sample-size performances of this method regarding the estimation of the parameters of the model as well as the measure of surrogacy. In Section 4, we apply this method to a meta-analysis of adjuvant chemotherapy for resectable gastric cancer to investigate time-to-relapse as a surrogate for overall survival. Finally, in Section 5, we conclude and provide some comments and further developments on the proposed methodology.

2. Mediation analysis and joint survival model

2.1. *Counterfactual framework for mediation analysis*

To give a causal interpretation of the "direct," "indirect," and "total" effects, we rely on the counterfactual results framework (Rubin, 1974; Holland, 1986). Based on this causal framework, Robins and Greenland (1992) gave conditions under which the identification of a direct and indirect effect of treatment by a post-treatment variable is possible, thus linking mediation analysis to the framework of causal inference. Much of the development of mediation analysis in the context of causal inference relies on the framework of potential outcomes, most commonly using the notion of nested counterfactual outcomes. Counterfactual mediation analysis for time-to-event outcomes was explored in Lin *and others* (2017) and in Tchetgen Tchetgen (2011).

Defining an indirect treatment effect in a setting where the intermediate outcome is a time-to-event that can be censored by the final event is challenging (Huang, 2020; Didelez, 2019). In this setting of semicompeting risks, Huang (2020) used counterfactual outcomes in a counting process to derive a nonparametric estimator of the indirect effect. In this article, we chose to use a flexible parametric model to estimate the direct and indirect effect of treatment on the final endpoint.

In the context of a meta-analysis, let ij denote the jth patient from the ith trial. Let Z_{ij} be a binary variable indicating whether the patient is in the experimental or control arm, X_{ij} a vector of baseline covariates, S_{ij} the uncensored surrogate (the mediator), and T_{ij} the uncensored final outcome. We assume an independent censoring C_{ij}. The observed censored outcomes are $S_{ij}^* = \min(S_{ij}, C_{ij}, T_{ij})$ and $T_{ij}^* = \min(T_{ij}, C_{ij})$ with their associated censoring indicators $d_{ij} = I_{S_{ij} \leq T_{ij}^*}$ and $\delta_{ij} = I_{T_{ij} \leq C_{ij}}$.

We define the counterfactual outcome $S_{ij}(z)$ as the counterfactual surrogate that would have been observed if a patient had been treated with treatment $Z_{ij} = z$. Furthermore, we define the counterfactual outcome $T_{ij}(z, s)$ as the final endpoint that would have been observed if treatment had been set to z and the surrogate had occurred at time s. We assume that for all $t < s$, the probability $\mathbb{P}\left(T_{ij}(z, s) \geq t\right)$ is independent of s, i.e., the occurrence of the surrogate at time s does not affect the distribution of $T_{ij}(z, s)$ before s. The nested counterfactual outcome $T_{ij}(z, S_{ij}(z'))$ is defined as the final endpoint that would have been observed if the treatment for the final endpoint had been set to z but the surrogate had "behaved" as if the treatment had been set at z'.

In the setting of semicompeting risks, such nested counterfactuals may be ill-defined, for example, if survival under z' is shorter than under z (Didelez, 2019). In this case, $T_{ij}(z, S_{ij}(z'))$ might not be properly defined if the support of $S_{ij}(z')$ is constrained by $T_{ij}(z', S_{ij}(z'))$.

We assume that:

1. $S_{ij}(z')$ can be censored by $T_{ij}(z, S_{ij}(z'))$
2. If $S_{ij}(z')$ is censored by $T_{ij}(z, S_{ij}(z'))$ it cannot have an effect on the occurrence of $T_{ij}(z, S_{ij}(z'))$
3. Thus $S_{ij}(z')$ can only have an effect on $T_{ij}(z, S_{ij}(z'))$ if it occurs first.

In the following, we will mainly focus on the hazard functions of $S_{ij}(z')$ and $T_{ij}(z, S_{ij}(z'))$.

Under the assumptions defined below, the observed outcome T_{ij} of a patient receiving treatment $Z_{ij} = z$ is equal to the counterfactual outcome $T_{ij}(z, S_{ij}(z))$. From these counterfactual outcomes, we can derive

a definition of the natural direct and indirect effects of treatment on T_{ij}. To identify the distributions of these counterfactual outcomes from the observed outcomes S_{ij} and T_{ij}, we must rely on a set of hypotheses (Imai *and others*, 2010).

In this counterfactual framework, we rely on the SUTVA (stable unit treatment value assumption). The SUTVA first assumes that each patient has only one outcome under each version of the treatment. This is plausible in settings where each version of the treatment is uniquely defined. The SUTVA also assumes that one's treatment assignment does not affect the outcome of another. This is also plausible in the setting of randomized trials except for some particular cases like in infectious diseases. In addition to the SUTVA, we must also rely on assumptions that are required in the context of mediation analysis.

First, we make the consistency assumption:

(A1) If a patient receives treatment $Z_{ij} = z$, then $S_{ij} = S_{ij}(z)$ and $T_{ij} = T_{ij}(z, S_{ij}(z))$.

This assumption states that a subject's observed outcomes S_{ij} and T_{ij} are the counterfactual outcomes associated with the treatment that the subject received.

We also make two positivity assumptions: let $f_{S_{ij}}(\cdot)$ be the density of S_{ij}, then we assume that,

(A2) $\forall z \in \{0, 1\}, \mathbb{P}(Z_{ij} = z \mid X_{ij}) > 0$
(A3) $\forall z, x, f_{S_{ij}}(s \mid Z_{ij} = z, X_{ij} = x) > 0, \forall s \geq 0$

Assumption (A2) states that there is no level of X that fully determines whether a patient is going to be treated or not. This assumption holds if Z is randomized. Assumption (A3) states that there is no level of (Z_{ij}, X_{ij}) for which a time of occurrence of S_{ij} would be impossible to observe: the support of the distribution of $S_{ij} \mid Z_{ij}, X_{ij}$ remains the same for all values of (Z_{ij}, X_{ij}).

Finally, the third assumption is the sequential ignorability assumption, which is also composed of two assumptions:

(A4) $\big(T_{ij}(z', s), S_{ij}(z)\big) \perp\!\!\!\perp Z_{ij} \mid X_{ij}, \forall z, z', s$
(A5) $T_{ij}(z', s) \perp\!\!\!\perp S_{ij}(z) \mid (Z_{ij} = z, X_{ij}), \forall z, z', s.$

Assumption A4 states that for patients in the same stratum of X_{ij}, the potential outcomes $\big(T_{ij}(z', s), S_{ij}(z)\big)$ are the same (in distribution) whether these patients are treated or not. This assumption usually holds when the treatment assignment is random. Assumption (A5) states that the potential outcomes $T_{ij}(z', s)$ are independent of $S_{ij}(z)$ for patients sharing the same baseline characteristics and receiving the same treatment $Z_{ij} = z$. This assumption holds under the assumption that there are no unmeasured confounders along the path from S_{ij} to T_{ij}, which is often a strong assumption even for randomized trials. However, as will be discussed in more details in the next section, the proposed approach relies on a parametric model that requires a less stringent assumption.

In the setting of mediation analysis, one must also rely on the so-called "cross-world" assumption (Andrews and Didelez, 2020) which in our case would be translated as follows:

(A6) $T_{ij}(z', s) \perp\!\!\!\perp S_{ij}(z), \forall s, z \neq z'.$

This is a strong assumption but we will see below that it can hold given the proposed statistical model.

We now define, for two treatment levels z and z' and for $t \geq 0$, the counterfactual survival function $\mathbb{S}_{z'z}(t)$ associated with the counterfactual outcome $T_{ij}(z', S_{ij}(z))$: $\mathbb{S}_{z'z}(t) = \mathbb{P}\left(T_{ij}(z', S_{ij}(z)) > t\right)$. From this function, the *natural indirect effect* (NIE) of Z_{ij} on T_{ij} through S_{ij} can be defined as $\text{NIE}(t) = \mathbb{S}_{11}(t) - \mathbb{S}_{10}(t)$.

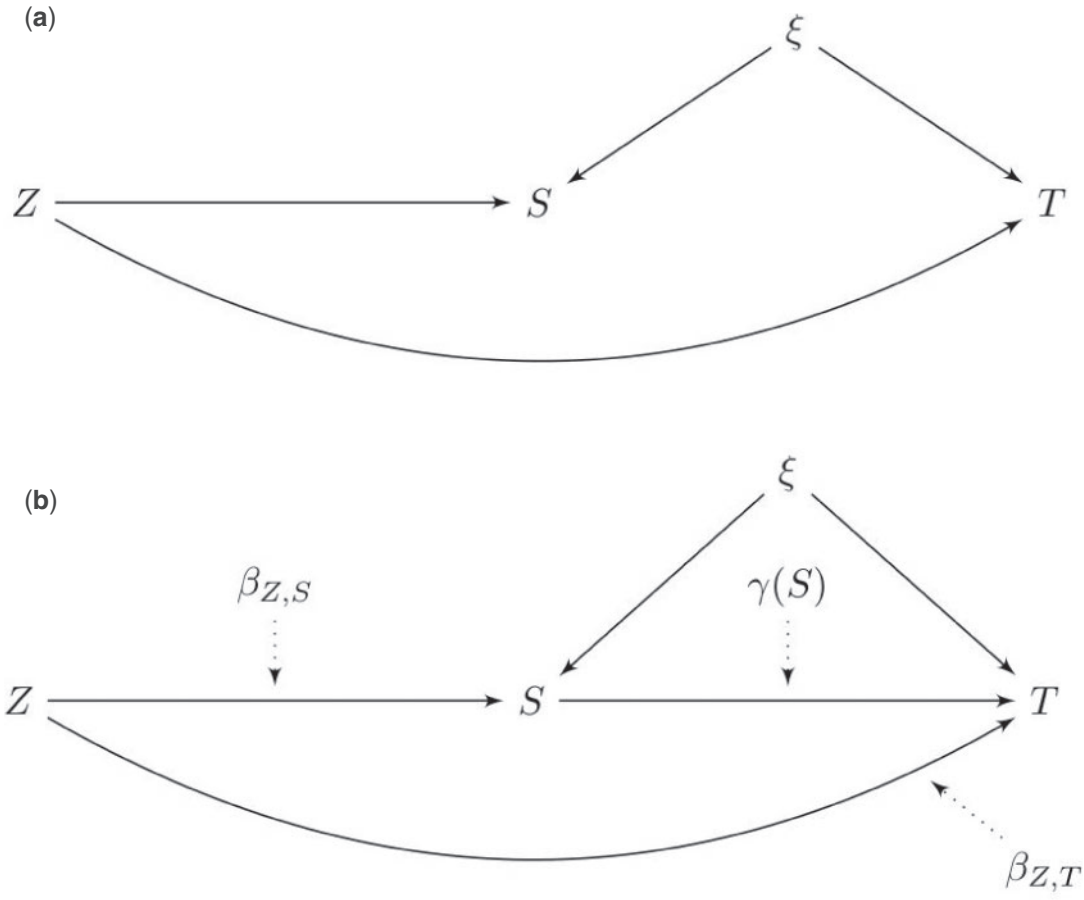

Fig. 1. (a) Joint model by Sofeu *and others* (2019): given the random effects (ξ), S and T are independent, and thus there is conditionally no direct effect of Z through S. (b) A generalization of (a) where S and T are not conditionally independent upon ξ but a direct link from S to T remains through the function $\gamma(S)$ as in (2.2).

The quantity $\text{NIE}(t)$ is the difference between the survival of a subject with treatment Z_{ij} set to $Z_{ij} = 1$ for both T_{ij} and S_{ij} and the survival of a subject with treatment set to $Z_{ij} = 1$ for T_{ij} but set to $Z_{ij} = 0$ for S_{ij}.

In the same way, the *natural direct effect* (NDE) is defined as $\text{NDE}(t) = \mathbb{S}_{10}(t) - \mathbb{S}_{00}(t)$ and the total effect (TE) as $\text{TE}(t) = \mathbb{S}_{11}(t) - \mathbb{S}_{00}(t)$. The proportion of treatment effect (PTE) is then defined as:

$$\text{PTE}(t) = \frac{\mathbb{S}_{11}(t) - \mathbb{S}_{10}(t)}{\mathbb{S}_{11}(t) - \mathbb{S}_{00}(t)} = \frac{\text{NIE}(t)}{\text{TE}(t)}. \tag{2.1}$$

The quantity $\text{PTE}(t)$ represents the proportion of the treatment effect on T_{ij} that is due to the effect of the treatment on S_{ij}. A high value for $\text{PTE}(t)$ means that a large fraction of the treatment effect on T is due to its effect on S_{ij}, making S_{ij} an interesting surrogate.

In the following section, we present a joint model for two time-to-event endpoints that also accounts for meta-analytic data, and we show how the functions $\mathbb{S}_{z'z}(t)$ (and hence $\text{PTE}(t)$) can be computed from this model.

2.2. *Frailty model for two time-to-event endpoints*

As in the previous section, T_{ij} and S_{ij} are the uncensored final and surrogate endpoint, respectively, and X_{ij} is a vector of baseline covariates. The joint distribution of (T_{ij}, S_{ij}) must be carefully described because the occurrence of T_{ij} may preclude the occurrence (or observation) of S_{ij}. Moreover, these endpoints are not independent at the individual level: subjects who are susceptible to shorter surrogate times S_{ij} may also be susceptible to shorter survival times T_{ij}. That is, some subjects are "more frail" than others, and a natural association between (T_{ij}, S_{ij}) at the individual level should be considered. Furthermore, we assume a meta-analytic setting in which the outcomes of subjects from the same cluster (a clinical center or a trial) are also associated. Therefore, two types of associations must be considered: at the individual and trial level. These two associations can be modeled with a joint model that uses shared random effects at each level. Sofeu *and others* (2019) have proposed such a model for validating surrogate endpoints using meta-analytic data. However, their model assumes independence between S_{ij} and T_{ij} conditionally on the vector of random effects ξ_{ij} and therefore no direct effect of S_{ij} on T_{ij} given ξ_{ij} is possible (see Figure 1(a)). For this model, $\text{NIE}(t \mid \xi_{ij}) = 0$ and thus $\text{PTE}(t \mid \xi_{ij}) = 0$. To allow a direct effect of S_{ij} on T_{ij} and to have a more comprehensive modeling, we decided to use the joint model given by Equation (2.2) below. In this model, risks of the final and surrogate endpoint are modeled assuming proportional hazards. A direct effect of S_{ij} on T_{ij} is enabled by a modification of the hazard function of T_{ij} once the surrogate endpoint S_{ij} occurs. This modification is accounted for by a function $\gamma(S_{ij})$ that depends on S_{ij}, see Figure 1(b). For an individual j from trial i, this model is defined as follows:

$$\begin{cases} \lambda_{S_{ij}(z)}(t) = \lambda_{0,S}(t) \exp\left(\omega_{ij} + u_i + z(\nu_{S,i} + \beta_{Z,S}) + \beta_S^{\mathsf{T}} X_{ij}\right) \\ \lambda_{T_{ij}(z,s)}(t) = \lambda_{0,T}(t) \exp\left(\zeta\omega_{ij} + \alpha u_i + z(\nu_{T,i} + \beta_{Z,T}) + \beta_T^{\mathsf{T}} X_{ij} + \gamma(s) I(s \le t)\right). \end{cases} \tag{2.2}$$

Here, ω_{ij} is an individual-level frailty, u_i a trial-level frailty representing the potential heterogeneity between patients across trials and $(\nu_{S,i}, \nu_{T,i})$ trial-level frailties allowing for a potential heterogeneity between treatment effects across trials. These frailties are assumed to be jointly Gaussian,

$$\begin{pmatrix} \omega_{ij} \\ u_i \\ \nu_{S,i} \\ \nu_{T,i} \end{pmatrix} \sim \mathcal{N}(\mu, \Sigma) \quad \text{where} \quad \mu = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \theta^2 & 0 & 0 & 0 \\ 0 & \rho^2 & 0 & 0 \\ 0 & 0 & \sigma_{\nu_S}^2 & \sigma_{\nu_{S,T}} \\ 0 & 0 & \sigma_{\nu_{S,T}} & \sigma_{\nu_T}^2 \end{pmatrix}. \tag{2.3}$$

The parameters (ζ, α) are power parameters that allow the frailties to affect hazard functions differently. The model parameters are estimated by maximizing a penalized likelihood,

$$L_{\text{pen}}(\phi) = L(\phi) - \kappa_T \int_0^\infty \left(\lambda_{0,T}''(t)\right)^2 \mathrm{d}t - \kappa_S \int_0^\infty \left(\lambda_{0,S}''(t)\right)^2 \mathrm{d}t, \tag{2.4}$$

where ϕ is the vector of all parameters of the model and $L(\phi)$ is the likelihood of the data. The parameters κ_T and κ_S are chosen by cross-validation and penalize the likelihood by the integral of the second derivative of the baseline hazards to ensure smooth estimates of these functions. Details on likelihood construction and estimation can be found in Appendix A of the Supplementary material at *Biostatistics* online.

In model given by Equation (2.2), a direct effect of $S_{ij}(z')$ on $T_{ij}(z, s)$ is introduced through the term $\gamma(s) I(s \le t)$. The indicator $I(s \le t)$ ensure that the risk of occurrence of $T_{ij}(z, s)$ only depends of s from the time of occurrence s.

The strength of an occurrence of $S_{ij}(z')$ on the risk of occurrence of $T_{ij}(z, s)$ is taken into account through the function $\gamma(s)$ with $\gamma(s) = 0$ indicating no effect.

The parameters $\beta_{Z,S}$ and $\beta_{Z,T}$ represents the direct effect of the treatment on $S_{ij}(z')$ and $T_{ij}(z,s)$, respectively. To have an indirect effect of the treatment on $T_{ij}(z,S_{ij}(z'))$ through $S_{ij}(z')$, both $\beta_{Z,S}$ and the function $\gamma(s)$ should be different from 0. Figure 5 in Appendix F of the Supplementary material at *Biostatistics* online shows an example of how the occurrence of $S_{ij}(z')$ can change the risk of occurrence of $T_{ij}(z,S_{ij}(z'))$.

Following model (2.2) and assuming that $T_{ij} = T_{ij}(z,S_{ij}(z))$ and $S_{ij} = S_{ij}(z)$, we have $T_{ij}(1,s) \perp\!\!\!\perp S_{ij}(0) \mid (X_{ij},\xi_{ij})$, where $\xi = (\omega,u,\nu_S,\nu_T)$ is the vector of random effects. Thus, the cross-world assumption that the counterfactual $T(z,s)$ is independent of $S(z')$ holds conditionally on X and ξ.

As pointed out in the previous section, this model requires a less stringent assumption than assumption (A5), since it does not require that (Z_{ij},X_{ij}) capture all the confounding between $S_{ij}(z')$ and $T_{ij}(z,S_{ij}(z'))$, but rather that this confounding should be captured by (Z_{ij},X_{ij},ξ_{ij}). Thus, we make a somewhat less stringent assumption:

(A5') $T_{ij}(z',s) \perp\!\!\!\perp S_{ij}(z) \mid (Z_{ij}=z,X_{ij},\xi_{ij}), \forall z,z',s$

Similarly, conditioning on X_{ij} in assumptions (A2)–(A4) is replaced in this model by a conditioning on (X_{ij},ξ_{ij}).

From (2.2), we can derive the hazard function for the observed data as:

$$\lambda_{S_{ij}}\left(t \mid \xi_{ij},Z_{ij},X_{ij}\right) = \lambda_{0,S}(t)\exp\left(\omega_{ij}+u_i+Z_{ij}(\nu_{S,i}+\beta_{Z,S})+\beta_S^{\mathsf{T}}X_{ij}\right)$$

and

$$\lambda_T\left(t \mid \xi_{ij},Z_{ij},S_{ij},X_{ij}\right) = \lambda_{0,T}(t)\exp\left(\zeta\omega_{ij}+\alpha u_i+Z_{ij}(\nu_{T,i}+\beta_{Z,T})+\beta_T^{\mathsf{T}}X_{ij}+\gamma(S_{ij})I\left(S_{ij}\le t\right)\right)$$

Under the assumptions (A1)–(A5') and (A6), we can write the "counterfactual" survival function $\mathbb{S}_{z'z}(t)$ as:

$$\mathbb{S}_{z'z}(t) = \int_{\mathcal{X}}\int_0^{\infty}\int_{\mathbb{R}^4}\mathcal{S}_T(t \mid s,z',\xi,x)f_S(s \mid z,\xi,x)f(\xi)f_X(x)\mathrm{d}s\mathrm{d}\xi\,\mathrm{d}x, \tag{2.5}$$

where the function $\mathcal{S}_T(t \mid s,z,\xi,x)$ is the conditional survival function of T obtained from $\lambda_T(t \mid \xi,z,s,x)$ and $f_S(s \mid z',\xi,x)$ is the conditional density function of S derived from $\lambda_S(t \mid \xi,z,x)$. The sets $\mathcal{X}$ and $\mathbb{R}^4$ are the supports of the covariates X and ξ, respectively. From the computation of $\mathbb{S}_{z'z}(t)$, we can derive $\mathrm{PTE}(t)$ according to (2.1) by considering appropriate values for z and z'.

A confidence interval for $\mathrm{PTE}(t)$ can be computed using either nonparametric or parametric bootstrap. The latter proceeds by taking the (estimated) covariance matrix of the maximum likelihood estimator (MLE) $\hat{\phi}$, $\hat{\Sigma}_{\hat{\phi}}$. Using the asymptotic normality of the MLE, we can generate a large number of vector of parameters $\phi_i \sim \mathcal{N}\left(0,\hat{\Sigma}_{\hat{\phi}}\right)$. For each simulated ϕ, we compute a function $\mathrm{PTE}_i(t)$. The empirical standard deviation of the vector $(\mathrm{PTE}_i(t))$ can then be used to computed a confidence interval for $\mathrm{PTE}(t)$ as $\mathrm{PTE}(t) \pm 1.96 \times \mathrm{se}$.

Furthermore, as described in Sofeu *and others* (2019), this model can be used to calculate the coefficient of determination of the trial-level random effects, R^2_{trial}, defined as $\frac{\sigma^2_{\nu_{S,T}}}{\sigma^2_{\nu_S}\sigma^2_{\nu_T}}$ and the Kendall's τ which measures the individual-level association between S and T.

3. Simulations

We conducted a simulation study to assess the performance of this model and the associated estimation procedure. We simulated scenarios with different numbers of trials, different variances for the random

effects ω and u and scenarios in which there was an indirect treatment effect on the final endpoint through the surrogate and scenarios in which there was no indirect effect.

3.1. *Simulation scheme*

Observations were generated from (2.2), where T was generated using a piecewise-exponential distribution.

We first generated the frailties according to (2.3) and the surrogate S according to the model (2.2) with a Weibull baseline hazard function for $\lambda_{0,S}(t)$. Then, a piecewise exponential model is applied to T, which consists of two parts. First, on $[0, S]$, the hazard of T conditionally on the random effects is given as in (2.2) with a constant baseline hazard. Then, on $[S, \infty[$, the hazard is simply that of the first interval multiplied by $e^{\gamma(S)}$. The procedure for generating T is detailed in Appendix B of the Supplementary material at *Biostatistics* online. After generating S and T, we applied a constant censoring and retrieved the associated censored follow-up times and event indicators. This resulted in a censoring rate of about 40% for S and T. The simulation scheme is fully detailed in Appendix C of the Supplementary material at *Biostatistics* online.

Several scenarios were considered for which the number of trials, K, was set to 20 or 50 and the function $\gamma(s)$ was equal to 0 or $\gamma(s) = e^{-cs^2}$ for $c = 0.001$. For the case where $\gamma(s) = 0$, there is no direct effect of S on T and thus no indirect effect of Z on T through S. In this case, by definition, we have $\text{PTE}(t) = 0$. The case where $\gamma(s) = e^{-cs^2}$ defines a small mediation where $\text{PTE}(t)$ is approximately equal to 0.15 at maximum follow-up. This function was chosen to yield a positive direct effect of S on T that smoothly decreases towards 0 over the time scale of the simulations.

We also examined scenarios with different values for the variances θ^2 and ρ^2 of the random effects ω_{ij} and u_i respectively. A first scenario with high variances where $\theta^2 = \rho^2 = 1$ and a second with lower variances where $\theta^2 = \rho^2 = 0.64$. Patients were randomly assigned to one of the Kth trials with equal probability K^{-1}. All patients in the same trial had the same frailties $(u_i, \nu_{S,i}, \nu_{T,i})$ to create a cluster effect across trials. For all these scenarios, the total number of patients was set at 1500. Each scenario was repeated 300 times.

To determine the true value of $\text{PTE}(t)$ for each scenario, we simulated data sets in which treatment was fixed for each endpoint. In this way, we were able to use a nonparametric estimate of the values of the survival functions $\mathbb{S}_{z'z}(t)$ associated with the data set in which the treatment values for S and T were fixed at z and z', respectively. We evaluated $\text{PTE}(t)$ for t ranging between 10 and 45 with a step of 1. These calculations were used in the results to compute the true value of $\text{PTE}(t)$ at three time points (20, 30 and 40) and in the graphical representation of the estimated $\text{PTE}(t)$ versus to the true $\text{PTE}(t)$.

3.2. *Results*

Estimations were evaluated for each parameter of the model in terms of bias, mean standard error, empirical standard deviation, and 95% coverage rate. For $\text{PTE}(t)$, we focused on three specific time points (20, 30, and 40). We also evaluated the estimations of the NIE and NDE at these time points as well as estimations of R^2_{trial} and the Kendall's τ.

Simulations results are shown in Table 1 for the cases with $\theta^2 = \rho^2 = 1$ and in Table 2 for the cases with $\theta^2 = \rho^2 = 0.64$. In Figure 2, we plotted the estimated $\text{PTE}(t)$ from each simulated data set with their average at each time compared to the true value of the function $\text{PTE}(t)$.

From Tables 1 and 2, it can be seen that increasing the number of trials improves the estimation of the parameters in terms of bias and coverage rate. This difference mainly affects the variances of the trial-level random effects (parameters $\sigma^2_{\nu_S}$, $\sigma^2_{\nu_T}$, and $\sigma_{\nu_{S,T}}$) and the parameters $\beta_{Z,S}$ and $\beta_{Z,T}$ of the treatment effects on S and T.

Table 1. *Simulations results for scenarios where* $\theta^2 = \rho^2 = 1$

		K=20 trials				K= 50 trials			
Parameter	Value	Est.†	SE.‡	SD.§	CR.¶	Est.	SE.	SD.	CR.
				$\gamma(s) = 0$					
PTE(20)	0.000	−0.000	0.214	0.138	0.976	0.004	0.116	0.107	0.957
PTE(30)	0.000	−0.002	0.186	0.157	0.945	−0.004	0.115	0.114	0.955
PTE(40)	0.000	−0.007	0.171	0.159	0.961	−0.017	0.115	0.108	0.965
NIE(20)	0.000	0.001	0.019	0.018	0.963	0.002	0.018	0.017	0.935
NIE(30)	0.000	0.002	0.021	0.022	0.940	0.000	0.020	0.020	0.945
NIE(40)	0.000	0.001	0.022	0.022	0.958	−0.002	0.022	0.020	0.967
NDE(20)	0.121	0.164	0.029	0.041	0.594	0.164	0.025	0.032	0.562
NDE(30)	0.184	0.181	0.031	0.045	0.812	0.188	0.027	0.035	0.889
NDE(40)	0.195	0.190	0.032	0.048	0.835	0.204	0.029	0.036	0.846
$\beta_{Z,S}$	−1.000	−0.760	0.211	0.348	0.673	−1.029	0.196	0.194	0.950
$\beta_{Z,T}$	−1.000	−1.046	0.209	0.264	0.929	−1.106	0.209	0.208	0.957
θ^2	1.000	1.096	0.507	0.498	0.929	1.065	0.550	0.584	0.887
ρ^2	1.000	1.832	0.496	1.323	0.673	1.096	0.357	0.402	0.935
$\sigma^2_{\nu_S}$	0.700	1.107	0.368	0.598	0.780	0.803	0.305	0.323	0.947
$\sigma^2_{\nu_T}$	0.700	0.682	0.307	0.434	0.775	0.819	0.337	0.384	0.930
$\sigma_{\nu_{S,T}}$	0.420	0.485	0.229	0.348	0.838	0.512	0.224	0.260	0.935
ζ	1.000	1.067	0.397	0.498	0.919	1.160	0.488	0.575	0.932
α	1.000	1.003	0.167	0.206	0.911	1.052	0.199	0.199	0.957
τ	0.395	0.416	—	0.072	0.757	0.370	—	0.051	0.935
R^2_{trial}	0.360	0.396	0.151	0.207	0.817	0.430	0.164	0.168	0.910
				$\gamma(s) > 0$					
PTE(20)	0.081	0.136	0.235	0.168	0.933	0.141	0.119	0.118	0.929
PTE(30)	0.130	0.147	0.201	0.281	0.954	0.144	0.118	0.118	0.948
PTE(40)	0.133	0.172	0.182	0.213	0.942	0.150	0.118	0.122	0.959
NIE(20)	0.011	0.025	0.022	0.023	0.896	0.027	0.022	0.022	0.904
NIE(30)	0.027	0.030	0.024	0.024	0.948	0.030	0.024	0.024	0.953
NIE(40)	0.029	0.032	0.025	0.026	0.933	0.032	0.025	0.026	0.945
NDE(20)	0.123	0.146	0.027	0.040	0.764	0.158	0.027	0.031	0.721
NDE(30)	0.182	0.157	0.029	0.041	0.782	0.171	0.029	0.034	0.882
NDE(40)	0.189	0.160	0.030	0.044	0.758	0.176	0.030	0.034	0.901
$\beta_{Z,S}$	−1.000	−0.971	0.209	0.291	0.870	−0.977	0.181	0.178	0.948
$\beta_{Z,T}$	−1.000	−0.951	0.195	0.258	0.882	−0.988	0.177	0.172	0.942
θ^2	1.000	1.043	0.447	0.436	0.934	1.025	0.472	0.501	0.901
ρ^2	1.000	1.444	0.398	0.634	0.804	1.231	0.362	0.412	0.942
$\sigma^2_{\nu_S}$	0.700	0.772	0.290	0.453	0.870	0.806	0.295	0.396	0.904
$\sigma^2_{\nu_T}$	0.700	0.595	0.254	0.556	0.565	0.783	0.283	0.305	0.934
$\sigma_{\nu_{S,T}}$	0.420	0.423	0.196	0.387	0.737	0.545	0.216	0.257	0.929
ζ	1.000	1.068	0.341	0.360	0.952	1.071	0.356	0.355	0.951
α	1.000	1.055	0.155	0.167	0.964	1.017	0.151	0.147	0.962
τ	0.395	0.411	—	0.051	0.861	0.387	—	0.046	0.926
R^2_{trial}	0.360	0.473	0.174	0.215	0.831	0.486	0.157	0.155	0.888

†Mean of the estimated parameters.
‡Mean of the estimated standard errors.
§Empirical standard deviation.
¶Coverage rate of the 95% confidence interval.

Table 2. *Simulations results for scenarios where* $\theta^2 = \rho^2 = 0.64$

Parameter	Value	K=20 trials Est.†	SE.‡	SD.§	CR.¶	K=50 trials Est.	SE.	SD.	CR.
				$\gamma(s) = 0$					
PTE(20)	0.000	−0.032	0.333	0.588	0.939	−0.001	0.110	0.107	0.949
PTE(30)	0.000	0.100	0.458	1.585	0.952	0.007	0.109	0.099	0.968
PTE(40)	0.000	−0.016	0.266	0.138	0.963	0.003	0.109	0.102	0.971
NIE(20)	0.000	−0.000	0.017	0.018	0.939	0.001	0.017	0.017	0.946
NIE(30)	0.000	0.001	0.019	0.019	0.939	0.002	0.020	0.019	0.952
NIE(40)	0.000	−0.002	0.020	0.020	0.959	0.001	0.021	0.021	0.965
NDE(20)	0.121	0.180	0.027	0.063	0.371	0.170	0.026	0.030	0.507
NDE(30)	0.184	0.205	0.030	0.069	0.531	0.193	0.028	0.032	0.898
NDE(40)	0.195	0.223	0.031	0.073	0.517	0.211	0.030	0.036	0.861
$\beta_{Z,S}$	−1.000	−0.894	0.190	0.312	0.762	−0.988	0.167	0.154	0.952
$\beta_{Z,T}$	−1.000	−1.084	0.204	0.380	0.715	−1.025	0.182	0.193	0.958
θ^2	0.640	0.644	0.338	0.330	0.919	0.668	0.363	0.392	0.899
ρ^2	0.640	0.887	0.256	0.513	0.795	0.855	0.255	0.317	0.928
$\sigma^2_{\nu_S}$	0.700	0.890	0.310	0.437	0.859	0.623	0.229	0.258	0.817
$\sigma^2_{\nu_T}$	0.700	0.769	0.302	0.546	0.725	0.812	0.291	0.345	0.910
$\sigma_{\nu_{S,T}}$	0.420	0.395	0.203	0.442	0.634	0.388	0.180	0.215	0.857
ζ	1.000	1.170	0.519	0.635	0.930	1.189	0.541	0.723	0.920
α	1.000	1.051	0.171	0.203	0.950	1.043	0.173	0.211	0.942
τ	0.315	0.312	—	0.064	0.862	0.309	—	0.052	0.920
R^2_{trial}	0.360	0.335	0.132	0.233	0.708	0.346	0.146	0.174	0.891
				$\gamma(s) > 0$					
PTE(20)	0.081	0.063	0.303	1.865	0.874	0.119	0.099	0.107	0.918
PTE(30)	0.130	0.182	0.179	0.125	0.943	0.133	0.097	0.094	0.952
PTE(40)	0.133	0.171	0.164	0.129	0.937	0.139	0.097	0.096	0.952
NIE(20)	0.011	0.033	0.022	0.025	0.821	0.025	0.020	0.022	0.879
NIE(30)	0.027	0.038	0.024	0.025	0.925	0.031	0.023	0.022	0.960
NIE(40)	0.029	0.037	0.025	0.026	0.937	0.034	0.024	0.023	0.949
NDE(20)	0.123	0.157	0.029	0.048	0.613	0.179	0.027	0.029	0.429
NDE(30)	0.182	0.169	0.030	0.047	0.777	0.200	0.029	0.032	0.873
NDE(40)	0.189	0.175	0.031	0.051	0.739	0.208	0.031	0.034	0.873
$\beta_{Z,S}$	−1.000	−1.022	0.185	0.376	0.634	−0.979	0.175	0.172	0.944
$\beta_{Z,T}$	−1.000	−0.899	0.195	0.275	0.800	−1.052	0.184	0.158	0.977
θ^2	0.640	0.667	0.323	0.365	0.922	0.677	0.343	0.325	0.924
ρ^2	0.640	1.006	0.276	0.502	0.722	0.767	0.230	0.268	0.929
$\sigma^2_{\nu_S}$	0.700	0.756	0.264	0.529	0.694	0.803	0.270	0.298	0.935
$\sigma^2_{\nu_T}$	0.700	0.960	0.335	0.544	0.831	0.874	0.294	0.329	0.938
$\sigma_{\nu_{S,T}}$	0.420	0.530	0.200	0.421	0.694	0.553	0.209	0.265	0.881
ζ	1.000	1.095	0.396	0.469	0.934	1.116	0.428	0.469	0.932
α	1.000	1.036	0.144	0.167	0.922	1.048	0.158	0.150	0.966
τ	0.315	0.332	—	0.062	0.800	0.311	—	0.051	0.912
R^2_{trial}	0.360	0.422	0.121	0.232	0.675	0.448	0.141	0.176	0.805

†Mean of the estimated parameters.
‡Mean of the estimated standard errors.
§Empirical standard deviation.
¶Coverage rate of the 95% confidence interval.

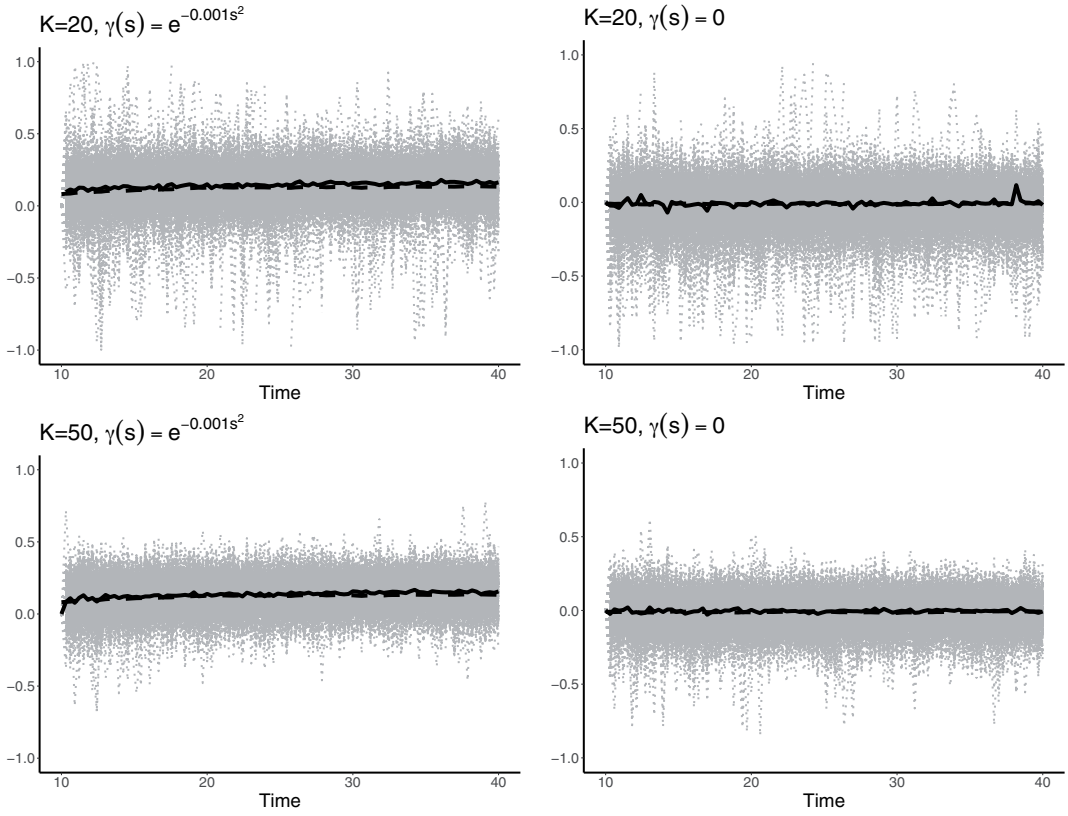

Fig. 2. Estimated PTE(t) for scenarios where $\theta^2 = \rho^2 = 1$ (dotted gray curves), their average (solid black curve), and the true curve (black dashed curve).

For trial-level random effects variances, the estimations are biased upward while their associated standard errors are biased downward, but increasing the number of trials reduces these biases. This bias is also reflected in the estimation of R^2_{trial} which is biased and results in a low coverage rate.

In addition, increasing the number of trials reduces the differences between the estimated standard errors of the different parameters and their empirical standard deviation (columns SE and SD respectively, in the tables). This improvement leads to a better coverage rate of the confidence interval for these parameters.

As for the estimation of the function PTE(t) and the NDE and NIE, from Tables 1 and 2, we see that both the NIE and PTE(t) as well as their standard errors are well estimated. Their confidence intervals also have coverage rate close to the nominal 95%. However, the estimation of natural direct effect (NDE) is biased which results in a low coverage rate. For the NDE, increasing the number of centers seems to moderately improve its estimation and the coverage rate, except for earlier timepoints ($t = 20$) where there is no clear improvement.

Figure 2 shows the estimations of PTE(t) for the cases where $\theta^2 = \rho^2 = 1$. Results for $\theta^2 = \rho^2 = 0.64$ are given in Figure 4 in Appendix F of the Supplementary material at *Biostatistics* online. These figures show the 300 estimated PTE(t) (one per simulated data set) in the dotted curves, their average at each time (solid black curve), and the true PTE(t) (dashed black curve). From these figures, we see that increasing the number of trials (K) from 20 (top panel) to 50 (bottom panel) reduces both the bias of the estimations of PTE(t) and its variance.

These figures also show that the bias of the estimator of PTE(t) is smaller for larger values of t. This could be due to the fact that for shorter times t the difference of survivals $\mathbb{S}_{11}(t) - \mathbb{S}_{00}(t)$ at the denominator of PTE(t) is likely close to zero (since both survivals are close to one) resulting in a less stable estimation of PTE(t).

This suggests that the estimation of PTE(t) should only be done for sufficiently large values of t. As can be seen in Figure 2, we only plotted the estimate of PTE(t) for $t \geq 10$ since the results for $t < 10$ were not relevant due to the very large instability of the estimations of PTE(t). This also suggests that before computing PTE(t), one should plot the function $\mathbb{S}_{11}(t)$ and $\mathbb{S}_{00}(t)$ to see if there are some time points where the two curves cross or are too close to avoid computing the function PTE(t) for those time points.

In all of the above scenarios, simulations included unmeasured confounders (simulated by the individual random effects) that were well specified with respect to the model: both the confounders at the simulation and at the estimation stages are Gaussian. We investigated scenarios in which these confounders were misspecified. First, with uniformly distributed confounders over [0, 1] and second, with binary confounders. The results of these additional simulations can be found in Appendix D of the Supplementary material at *Biostatistics* online. These results suggest that the approach is quite robust when there are (at least time-fixed) unmeasured confounders since the effects on the bias in estimating PTE(t) are limited. Moreover, we also investigated a case with a smaller number of trials ($K = 10$) whose results are given in Table 3. From these results, we can see that the method performs poorer with a lower number of centers, especially regarding the estimation of the standard errors of the parameters of the model as well as for PTE(t), suggesting that this method should be applied on studies with a sufficient number of trials/clusters.

4. APPLICATION

We applied this approach to the meta-analysis gastric comparing adjuvant chemotherapy with surgery alone in patients with resectable gastric cancer (Paoletti *and others*, 2010). We also used this method in a second application with data from a trial in second-line treatment of metastatic colorectal cancer (Peeters *and others*, 2014). The analysis and results of this second application can be found in Appendix F of the Supplementary material at *Biostatistics* online.

4.1. *Time-to-relapse as a surrogate of overall survival in gastric cancer*

We studied time-to-relapse (TTR) as a surrogate for overall survival (OS). OS was defined as the time from randomization to death from any cause, whereas TTR was defined as the time from randomization to disease recurrence or the occurrence of second cancer, whichever occurred first.

4.1.1. *Data description* The data set contains information on 3288 patients from 14 clinical trials. Of these 3288 patients, 1634 (49%) received adjuvant chemotherapy before surgery, and the remaining 1654 (51%) underwent surgery without prior adjuvant chemotherapy. The total number of deaths was 1705 (51.8%) and ranged from 13.4% to 84% in the 14 trials. A total of 1352 TTR events (41%) were observed, ranging from 10% to 73% across the trials. The median follow-up time was 4.78 years in the adjuvant chemotherapy group (range = [0.005, 24.88]) and 4.40 years in the control group (range = [0.005, 23.96]). Kaplan–Meier estimates of survival functions for both endpoints are provided in Figure 6 in Appendix G of the Supplementary material at *Biostatistics* online.

4.1.2. *Data analysis and results* We applied the model defined by (2.2) to this data set. The baseline hazard functions, $\lambda_{0,T}(t)$ and $\lambda_{0,S}(t)$, were estimated using splines with seven knots. For the estimation of the function $\gamma(s)$, cubic B-splines with one interior knot at the median of the observed surrogate times S

Table 3. *Simulations results with* $K = 10$ *centers and* $\gamma(s) > 0$

Parameters	Value	Est.†	SE.‡	SD.§	CR.¶
PTE(20)	0.081	0.180	0.719	0.194	0.970
PTE(30)	0.130	0.149	0.471	0.212	0.960
PTE(40)	0.133	0.120	0.698	0.398	0.980
NIE(20)	0.011	0.032	0.025	0.026	0.880
NIE(30)	0.027	0.036	0.028	0.029	0.900
NIE(40)	0.029	0.038	0.029	0.030	0.940
NDE(20)	0.123	0.160	0.036	0.065	0.550
NDE(30)	0.182	0.167	0.039	0.065	0.770
NDE(40)	0.189	0.173	0.040	0.068	0.740
$\beta_{Z,S}$	−1.000	−0.973	0.283	0.426	0.798
$\beta_{Z,T}$	−1.000	−0.939	0.286	0.383	0.862
θ^2	1.000	1.048	0.535	0.667	0.898
ρ^2	1.000	1.087	0.422	1.074	0.615
$\sigma^2_{\nu_S}$	0.700	0.783	0.435	0.560	0.798
$\sigma^2_{\nu_T}$	0.700	0.690	0.414	0.583	0.734
$\sigma_{\nu_{S,T}}$	0.420	0.436	0.303	0.421	0.817
ζ	1.000	1.154	0.498	0.564	0.881
α	1.000	1.047	0.232	0.274	0.899
τ	0.395	0.340	—	0.102	0.780
R^2_{trial}	0.360	0.515	0.197	0.239	0.872

†Mean of the estimated parameters.
‡Mean of the estimated standard errors.
§Empirical standard deviation.
¶Coverage rate of the 95% confidence interval.

were used. The results for the estimation of the main parameters of the model are given in Table 4. In this table, the estimations of PTE(t) at three time points (2, 3, and 8 years) are also reported. The graphical representation of the estimation of $\gamma(s)$ is given in Figure 3 and the estimated function PTE(t) is given in Figure 4(a). The estimated natural direct, indirect, and total effects associated with PTE(t) are shown in Figure 4(b).

From Table 4, we can see that there is a significant treatment effect on the surrogate ($\beta_{Z,S} = -0.47$), suggesting that adjuvant chemotherapy prolongs TTR. The estimation of $\beta_{Z,T}$, which is not significantly different from 0, suggests that the treatment has no direct effect on overall survival (see Figure 1(b)). In addition, the variances of the random effects ω_{ij} and u_i (θ^2 and ρ^2, respectively), which are significantly different from zero, suggest an association between S and T at the individual level and an heterogeneity among of the patients between the trials. This suggests that these frailties capture some of the effects of confounding factors that were not measured. However, the nonsignificant variances of ν_S, ν_T, and $\nu_{S,T}$ do not indicate heterogeneity in treatment effects between trials.

Furthermore, as shown in Figure 3, the estimation of $\gamma(s)$ suggests that there is a strong effect of TTR on OS. The estimated function being positive shows that the occurrence of a TTR event increases the risk of occurrence of death. Because the function is increasing, a late occurrence of a TTR event has a greater effect on increasing the risk of OS than an earlier occurrence of a TTR event.

Overall, $\hat{\gamma}(s)$ and $\hat{\beta}_{Z,S}$ being significantly different from zero indicates that there is a mediated effect of the treatment on OS through TTR and $\hat{\beta}_{Z,T}$ being close to zero implies that most of the total effect is

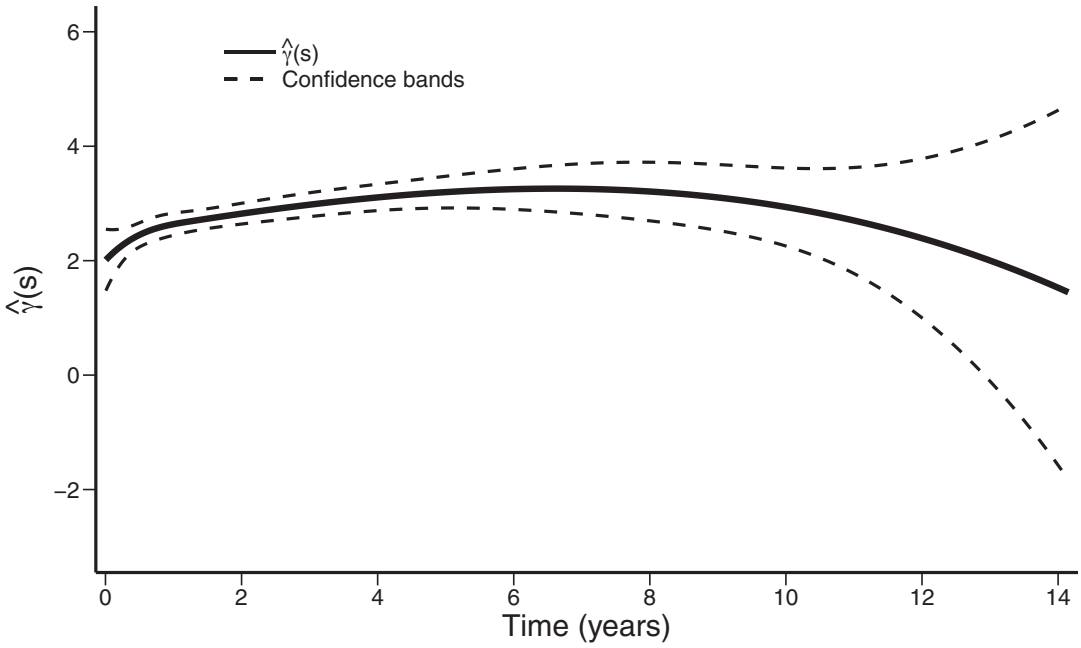

Fig. 3. Estimated $\hat{\gamma}(s)$ from the gastric application with its 95% pointwise confidence bands.

Table 4. *Estimated parameters of model (2.2) for the application on gastric cancer*

Parameter	Est.	Std. Err.	p-value
PTE(2)	0.853	0.283	—
PTE(3)	0.875	0.254	—
PTE(8)	0.839	0.155	—
$\beta_{Z,S}$	−0.453	0.104	<0.001
$\beta_{Z,T}$	−0.057	0.091	0.532
θ^2	3.807	0.317	<0.001
ρ^2	1.035	0.172	<0.001
$\sigma^2_{\nu_S}$	0.011	0.023	0.633
$\sigma^2_{\nu_T}$	0.018	0.027	0.495
$\sigma_{\nu_{S,T}}$	−0.014	0.011	0.212
ζ	0.681	0.043	<0.001
α	0.758	0.063	<0.001

this mediated effect. This can be seen in Figure 4(a), which shows that the estimated function $\hat{\text{PTE}}(t)$ is very close to 1.

Since $\hat{\beta}_{Z,S} < 0$ and $\hat{\gamma}(s) > 0$, patients treated with adjuvant chemotherapy have better survival, so their risk of death is likely to be increased at a later time point than patients in the control group, implying that the overall mediated treatment effect has a protective effect towards OS. This is shown in Figure 4(b), where the natural direct effect is not significantly different from zero over time, whereas both the natural indirect effect and the overall effect are significantly different from zero. From these results, it appears

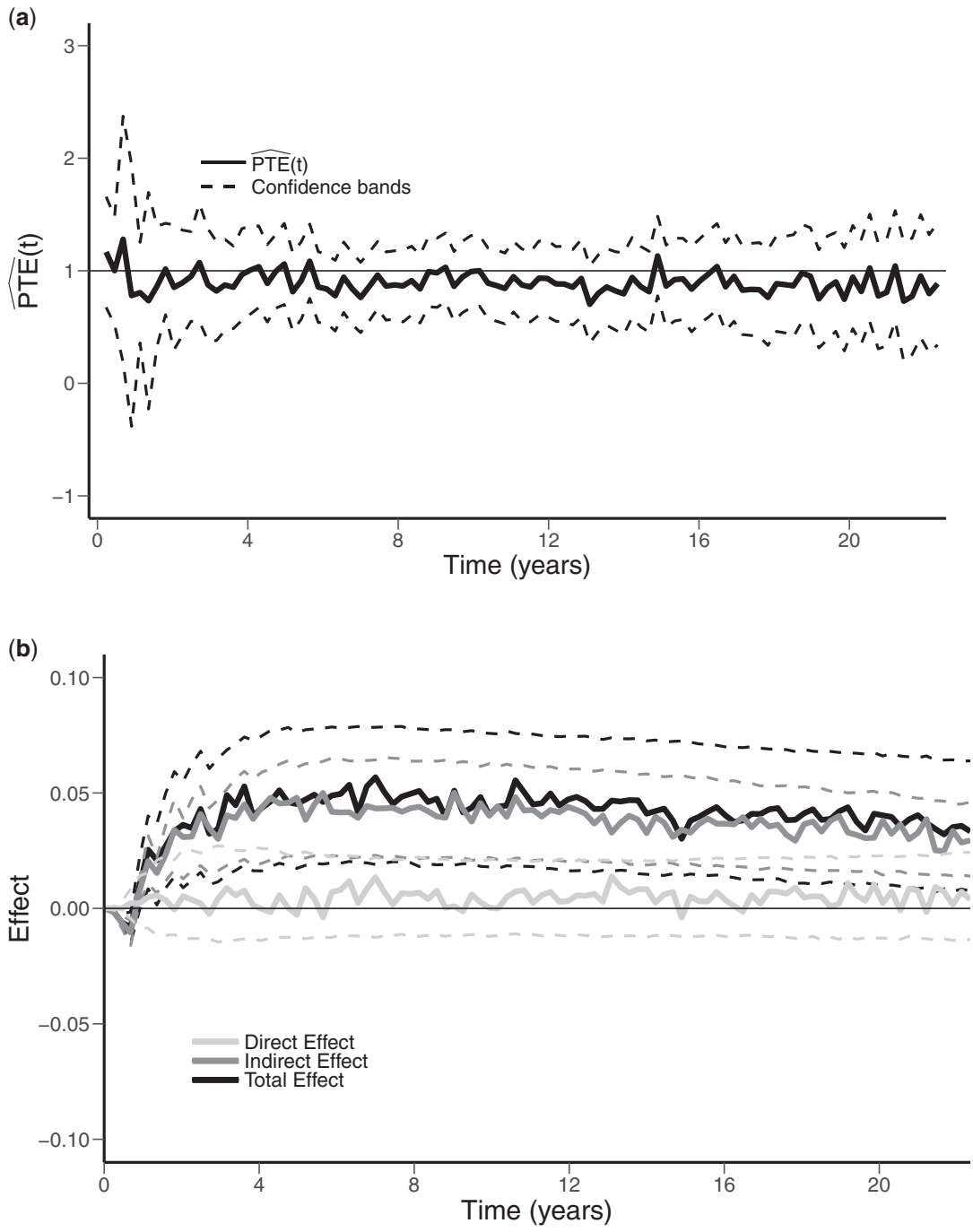

Fig. 4. Results from the application on gastric cancer. (a) Estimated PTE(t) and its 95% pointwise confidence bands. (b) Estimated natural effects and their 95% pointwise confidence bands.

that there is a significant overall effect of treatment on OS and that most of this effect comes from its effect on TTR. TTR is therefore likely to be a good surrogate for OS in the setting of adjuvant chemotherapy in patients with resectable gastric cancer.

These results had to be taken with caution as the limited number of trials (14) may cause bias in the estimated parameters of the model (and especially their standard errors). However, the high value of the estimated PTE(t) suggests that even if these results may be biased, it is likely that the true value PTE(t) is also high and that TTR is still a good surrogate for OS.

5. Discussion

In this article, we have proposed an approach to mediation analysis in the context of surrogate validation when both the surrogate and the final endpoint are time-to-event. This approach can be decomposed into two parts. First, using the counterfactual framework of causal inference, we defined a surrogacy measure that corresponds to the proportion of the treatment effect on the final endpoint that results from its effect on the surrogate. To calculate this proportion, we need to estimate the distribution of counterfactual outcomes, which leads us to the second part of this article, which deals with the statistical model used in this context.

To causally interpret the estimated direct and indirect effects of the treatment, this approach relies on several assumptions. In particular, assumption (A5') is crucial. To ensure that this assumption holds, the baseline covariates X and the random effects vector ξ must capture all of the confounding between S and T. It should be noted that there are situations where this is unlikely, for example, when there is time-varying confounding. A generalization of this method would include both time-varying covariates and time-varying random effects but our approach is currently limited to settings of time-fixed confounding.

Problems can arise with the function PTE(t) when both survival values appearing in the denominator are equal. This is the case, for example, when the survival functions $\mathbb{S}_{11}(t)$ and $\mathbb{S}_{00}(t)$ of the experimental and control groups cross. In this case, at this particular time t, there is no overall effect of treatment on the final endpoint, and PTE(t) is not defined. It may therefore be helpful to first examine the estimated Kaplan–Meier curves in each group to determine whether this is likely to be the case.

We used a counterfactual approach to define the survival functions $\mathbb{S}_{z'z}(t)$, which are then used to define the proportion PTE(t). To be able to identify these counterfactual survival functions from an observed data set, we made several assumptions. In particular, the sequential ignorability assumption warrants some attention and its plausibility might be investigated using sensitivity analyses.

The surrogacy measure proposed in this article is the function PTE(t). For clinical purposes, it may be more meaningful to assess surrogacy only at some relevant time points, for example, at 12 or 24 months. In this case, the function PTE(t) can be calculated only for these time points. However, the entire function PTE(t) is also of interest, as it can show a trend for surrogacy over time.

We would like to point out that while this approach to mediation analysis and the associated definition of indirect and direct effects is based on the counterfactual outcomes, other approaches were also proposed for time-to-event outcomes (Strohmaier *and others*, 2015; Aalen *and others*, 2020). These methods use dynamic modeling and do not use counterfactual outcomes to derive a causal definition of direct and indirect effects but are based on modeling structural equations using additive hazards in a continuous-time manner.

We proposed a method to validate a surrogate endpoint by studying the proportion of the treatment effect on the final endpoint that goes through the surrogate. The indirect treatment effect through the surrogate is composed of the direct treatment effect on the surrogate and the effect of the surrogate on the final endpoint. If at least one of these effects is absent, the indirect effect of treatment on the final endpoint is zero. However, as noted by VanderWeele (2013), in the case where there is no direct effect of the surrogate on the final endpoint, the surrogate could still be a good surrogate if, for example, there is a strong association between the surrogate and the final endpoint due to confounders. Model (2.2) can also

be used to study the association between the surrogate and the final endpoint through the R^2_{trial} and the Kendall's τ as described in Sofeu *and others* (2019). These quantities quantify the association between the two endpoints at the individual and trial levels. In particular, the R^2_{trial} is the coefficient of determination of the trial-level random effects. These quantities have been proposed for surrogate validation (Buyse *and others*, 2000). However, because they only measure association, these quantities do not provide information regarding the proportion of treatment mediated. Kendall's τ, which measures the individual-level association between S and T, would not be able to discriminate which proportion of this association is due to confounding or due to a direct link of S on T. Although a strong mediated effect may be sufficient for surrogacy, it is not a necessary condition for an endpoint to be a good surrogate. This model allows the use of both approaches, which complement each other, to evaluate a surrogate by either mediated effects or association.

As the simulations with $K = 10$ trials suggest, this method should be employed with a sufficient number of trials. However, other units of clustering could be used, such as centers of care for example.

Because the proposed approach focuses on the indirect effect of a treatment on a final endpoint through a surrogate, it may not be appropriate in cases where the surrogate is a composite endpoint that includes the final endpoint as an event. In such a case, the occurrence of the final endpoint leads to the occurrence of the surrogate and it is therefore difficult to define what would be a direct effect of the surrogate on the final endpoint since both endpoints would occur at exactly the same time. This is the case for progression-free survival when death is also accounted as a progression. The proposed approach is not appropriate for this situation, and it would be interesting to see if other approaches can be used to derive an appropriate definition of a direct and indirect effect for composite endpoints.

6. Software

The proposed method is available in the `jointSurroPenal` function of the R package `frailtypack` (Król *and others*, 2017) which is available on the R-CRAN at https://CRAN.R-project.org/package=frailtypack. The data set should have the following format:

```
head(data)
patientID    timeT    timeS statusT statusS trt trialID
        1 9.057946 2.217739       1       1   0       1
        2 2.986813 1.389263       1       1   0       1
        3 8.874237 8.874237       1       0   1       1
        4 3.245388 1.809671       1       1   1       1
        5 4.448964 2.603604       1       1   0       1
```

Where `timeT, timeS, statusT,` and `statusS` are respectively the time for the final and surrogate endpoints as well as their censoring indicators. The `R` command to call model (2.2) with one interior knot to estimate the function $\gamma(s)$ is:

```
require(frailtypack)
result<-jointSurroPenal(data,mediation=TRUE,g.nknots=1,
                        rt.ntimes=30,rt.nmc=10000,rt.boot=T,
                        rt.nboot=2000,rt.boot.nmc=10000)
```

After the execution of the program, summary of the estimations and related figures (Kaplan–Meier estimates, baseline hazards, $\gamma(s)$, PTE(t), and natural effects) can be obtained from the commands `summary(result)` and `plot(result)`.

Supplementary material

Supplementary material is available online at http://biostatistics.oxfordjournals.org.

Acknowledgments

Computations for this work were performed using HPC resources from the Mesocentre de Calcul Intensif Aquitain (MCIA). This publication is based on data obtained from www.projectdatasphere.org, which is maintained by Project Data Sphere. Neither Project Data Sphere nor the owner(s) of any information from the web site have contributed to, approved or are in any way responsible for the contents of this publication. The authors would like to thank the referees and associate editor whose comments greatly helped improve this manuscript.

Conflict of Interest: None declared.

Funding

This work was supported within the framework of the PIA3 (Investment for the future), project reference 17-EURE-0019.

References

Aalen, O. O., Stensrud, M. J., Didelez, V., Daniel, R., Røysland, K. and Strohmaier, S. (2020). Time-dependent mediators in survival analysis: modeling direct and indirect effects with the additive hazards model. *Biometrical Journal* **62**, 532–549.

Andrews, R. M. and Didelez, V. (2020). Insights into the cross-world independence assumption of causal mediation analysis. *Epidemiology* **32**, 209–219.

Buyse, M. and Molenberghs, G. (2005). *The Evaluation of Surrogate Endpoints*. (Vol. 427). T. Burzykowski (Ed.). New York: Springer.

Burzykowski, T., Buyse, M., Piccart-Gebhart, M. J., Sledge, G., Carmichael, J., Luck, H.-J., Mackey, J. R., Nabholtz, J.-M., Paridaens, R., Biganzoli, L. *and others*. (2008). Evaluation of tumor response, disease control, progression-free survival, and time to progression as potential surrogate end points in metastatic breast cancer. *Journal of Clinical Oncology* **26**, 1987–1992.

Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D. and Geys, H. (2000). The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics* **1**, 49–67.

Buyse, M., Molenberghs, G., Paoletti, X., Oba, K., Alonso, A., Van der Elst, W. and Burzykowski, T. (2016). Statistical evaluation of surrogate endpoints with examples from cancer clinical trials. *Biometrical Journal* **58**, 104–132.

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

Daniels, M. J. and Hughes, M. D. (1997). Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine* **16**, 1965–1982.

Didelez, V. (2019). Defining causal mediation with a longitudinal mediator and a survival outcome. *Lifetime Data Analysis* **25**, 593–610.

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

Freedman, L. S., Graubard, B. I. and Schatzkin, A. (1992). Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine* **11**, 167–178.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association* **81**, 945–960.

HUANG, Y.-T. (2020). Causal mediation of semicompeting risks. *Biometrics*.

IMAI, K., KEELE, L. AND YAMAMOTO, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science* **25**, 51–71.

JOFFE, M. M. AND GREENE, T. (2009). Related causal frameworks for surrogate outcomes. *Biometrics* **65**, 530–538.

KRÓL, A., MAUGUEN, A., MAZROUI, Y., LAURENT, A., MICHIELS, S. AND RONDEAU, V. (2017). Tutorial in joint modeling and prediction: a statistical software for correlated longitudinal outcomes, recurrent events and a terminal event. *Journal of Statistical Software* **81**, 1–52. https://doi.org/10.18637/jss.v081.i03

LEE, D.-W., JANG, M.-J., LEE, K.-H., CHO, E. J., LEE, J.-H., YU, S. J., KIM, Y. J., YOON, J.-H., KIM, T.-Y., HAN, S.-W. *and others*. (2016). TTP as a surrogate endpoint in advanced hepatocellular carcinoma treated with molecular targeted therapy: meta-analysis of randomised controlled trials. *British Journal of Cancer* **115**, 1201–1205.

LIN, S.-H., YOUNG, J. G., LOGAN, R. AND VANDERWEELE, T. J. (2017). Mediation analysis for a survival outcome with time-varying exposures, mediators, and confounders. *Statistics in Medicine* **36**, 4153–4166.

PAOLETTI, X., OBA, K., BURZYKOWSKI, T., MICHIELS, S., OHASHI, Y., PIGNON, J.-P., ROUGIER, P., SAKAMOTO, J., SARGENT, D., SASAKO, M. *and others*. (2010). Benefit of adjuvant chemotherapy for resectable gastric cancer: a meta-analysis. *JAMA* **303**, 1729–1737.

PEETERS, M., PRICE, T. J., CERVANTES, A., SOBRERO, A. F., DUCREUX, M., HOTKO, Y., ANDRÉ, T., CHAN, E., LORDICK, F., PUNT, C. J. A. *and others*. (2014). Final results from a randomized phase 3 study of folfiri±panitumumab for second-line treatment of metastatic colorectal cancer. *Annals of Oncology* **25**, 107–116.

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

ROBINS, J. M. AND GREENLAND, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, **3**, 143–155.

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 688.

SOFEU, C. L., EMURA, T. AND RONDEAU, V. (2019). One-step validation method for surrogate endpoints using data from multiple randomized cancer clinical trials with failure-time endpoints. *Statistics in Medicine* **38**, 2928–2942.

STROHMAIER, S., RØYSLAND, K., HOFF, R., BORGAN, Ø., PEDERSEN, T. R. AND AALEN, O. O. (2015). Dynamic path analysis—a useful tool to investigate mediation processes in clinical survival trials. *Statistics in Medicine* **34**, 3866–3887.

TCHETGEN TCHETGEN, E. J. (2011). On causal mediation analysis with a survival outcome. *The International Journal of Biostatistics* **7**, Article 33. DOI:10.2202/1557-4679.1351.

VANDENBERGHE, S., DUCHATEAU, L., SLAETS, L., BOGAERTS, J. AND VANSTEELANDT, S. (2018). Surrogate marker analysis in cancer clinical trials through time-to-event mediation techniques. *Statistical Methods in Medical Research* **27**, 3367–3385.

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

ZHENG, C. AND LIU, L. (2021). Quantifying direct and indirect effect for longitudinal mediator and survival outcome using joint modeling approach. *Biometrics* **78**, 1233–1243.

ZIGLER, C. M. AND BELIN, T. R. (2012). A Bayesian approach to improved estimation of causal effect predictiveness for a principal surrogate endpoint. *Biometrics* **68**, 922–932.

[*Received November 22, 2021; revised October 24, 2022; accepted for publication October 25, 2022*]