Evaluating time-to-event surrogates for time-to-event true endpoints: an information-theoretic approach based on causal inference

Peer-reviewed author version

Stijven, Florian; MOLENBERGHS, Geert; VAN KEILEGOM, Ingrid; VAN DER ELST, Wim & ALONSO ABAD, Ariel (2024) Evaluating time-to-event surrogates for time-to-event true endpoints: an information-theoretic approach based on causal inference. In: Lifetime data analysis,.

DOI: 10.1007/s10985-024-09638-7
Handle: http://hdl.handle.net/1942/44588

Lifetime Data Analysis manuscript No.
(will be inserted by the editor)

Evaluating Time-to-Event Surrogates for Time-to-Event True Endpoints: An Information-Theoretic Approach Based on Causal Inference

Florian Stijven · Geert Molenberghs · Ingrid Van Keilegom · Wim Van der Elst · Ariel Alonso

Received: date / Accepted: date

Abstract Putative surrogate endpoints must undergo a rigorous statistical evaluation before they can be used in clinical trials. Numerous frameworks have been introduced for this purpose. In this study, we extend the scope of the information-theoretic causal-inference approach to encompass scenarios where both outcomes are time-to-event endpoints, using the flexibility provided by D-vine copulas. We evaluate the quality of the putative surrogate using the individual causal association (ICA)—a measure based on the mutual information between the individual causal treatment effects. However, in spite of its appealing mathematical properties, the ICA may be ill defined for composite endpoints. Therefore, we also propose an alternative rank-based metric for assessing the ICA. Due to the fundamental problem of causal inference, the joint distribution of all potential outcomes is only partially identifiable and, consequently, the ICA cannot be estimated without strong unverifiable assumptions. This is addressed by a formal sensitivity analysis that is summarized by the so-called intervals of ignorance and uncertainty. The frequentist properties of these intervals are discussed in detail. Finally, the proposed methods are illustrated with an analysis of pooled data from two advanced colorectal cancer trials. The newly developed techniques have been implemented in the R package *Surrogate*.

F. Stijven
KU Leuven, I-BioStat, Leuven, Belgium
E-mail: florian.stijven@kuleuven.be

G. Molenberghs
KU Leuven, I-BioStat, Leuven, Belgium
Universiteit Hasselt, I-BioStat, Hasselt, Belgium

I. Van Keilegom
KU Leuven, ORSTAT, Leuven, Belgium

Wim Van der Elst
The Janssen Pharmaceutical companies of Johnson & Johnson, Beerse, Belgium

Ariel Alonso
KU Leuven, I-BioStat, Leuven, Belgium

Keywords Individual causal association · Sensitivity analysis · Surrogates · Survival analysis · Vine copula

1 Introduction

Drug development is a long, complex, and costly process, characterized by a substantial degree of uncertainty about the eventual success of the drug (Pankevich et al., 2014). A major challenge in this process is the selection of the primary endpoint. This is, ideally, the clinically most relevant endpoint, which is further termed the *true endpoint* and denoted by T (Alonso et al., 2017). In practice, true endpoints often present significant limitations, including the need for long follow-up (e.g., overall survival (OS) in early cancer stages) or costly measurements (e.g., imaging data). A seemingly attractive strategy to tackle these challenges involves replacing the true endpoint with an alternative outcome where the aforementioned issues are less prominent. However, not just any alternative endpoint can serve as a suitable replacement; it requires a rigorous statistical evaluation. Upon successful evaluation, the alternative endpoint is recognized as a valid *surrogate endpoint*. Throughout the manuscript, the potential surrogate endpoint is denoted by S.

In the early days, a prevalent misconception was that a robust association between the putative surrogate and the true endpoint alone would suffice for the statistical evaluation of the former. This misconception led to unfortunate events, prompting the development of more sophisticated evaluation methods (Alonso et al., 2017, Section 1.1.3). The first formal statistical framework for evaluating surrogate endpoints was introduced by Prentice (1989) in the single-trial setting (STS) where data on the surrogate and the true endpoint are available from a single clinical trial. Although Prentice's proposal was pivotal for the assessment of surrogate endpoints, it is currently recognized as inadequate for various reasons (Buyse et al., 2000a).

Buyse et al. (2000b) introduced the meta-analytic framework, which is based on meta analysis and separates trial-level from individual-level surrogacy. The meta-analytic framework is nowadays considered the gold standard for evaluating surrogate endpoints (Alonso et al., 2017). However, this framework relies on data from multiple trials. Such data are often unavailable in the early stages of drug development, precisely the moment when surrogate endpoints are most crucial. Consequently, the advancement of methods for the STS remains important.

In the STS, Alonso et al. (2015) introduced a new framework based on information-theoretic and causal-inference concepts. This approach is known as the information-theoretic causal-inference (ITCI) framework. It is based on the individual causal association (ICA), a metric of surrogacy that quantifies the association between individual causal treatment effects on the surrogate and the true endpoint. The ICA is grounded in information theory and offers a coherent and intuitive interpretation across various types of endpoints. Estimating the ICA requires a model for the four-dimensional vector of potential outcomes associated with the true and surrogate endpoints. This model is only partially identifiable and, hence, the ICA cannot be consistently estimated from the available data. To address this issue, Alonso et al. (2015) proposed a sensitivity analysis. Methods in the ITCI framework have been developed only for binary

and continuous endpoints (Alonso et al., 2015; Meyvisch et al., 2020; Alonso Abad et al., 2024), and the inferential framework used in the proposed sensitivity analysis has yet to be formalized.

In certain clinical domains, such as oncology, the true and putative surrogate endpoints are often time-to-event endpoints; this is the setting to which we extend the ITCI framework in this paper. This extension relies on a flexible D-vine copula model for the vector of potential outcomes. This model can handle a diverse range of non-normally distributed surrogate and true endpoints and can account for right-censoring. Furthermore, we formalize the inferential framework underlying the sensitivity analysis in the ITCI framework. Finally, we apply the new approach to data from two trials in advanced colorectal cancer, where both endpoints are time-to-event variables with semi-competing risks.

In Section 2, we introduce the advanced colorectal cancer data and corresponding notation. In Section 3, we summarize the ITCI framework. We introduce the D-vine copula model in Section 4. In Section 5, we formalize the sensitivity analysis and describe how it can be implemented in practice. Next, we illustrate our approach with an analysis of the advanced colorectal cancer data in Section 6, and we end with some concluding remarks.

2 Case Study and Notation

The advanced colorectal cancer data are pooled data from two randomized trials (Corfu-A Study Group, 1995; Greco et al., 1996) with 496 and 245 patients, respectively. In these trials, patients were randomized with equal allocation to receive (i) 5FU plus interferon (control treatment, $A = 0$), or (ii) 5FU alone or with folinic acid (experimental treatment, $A = 1$) for the treatment of advanced colorectal cancer. These trials found no significant treatment effects on progression-free survival (PFS, defined in the next paragraph) or overall survival (OS). In the combined data, the median PFS and OS times are, respectively, 25.4 weeks (95% CI, 24.1 to 27.9) and 52.3 weeks (95% CI, 49.3 to 57.1); the corresponding censoring rates are 5.7% and 25.0 %. Independent censoring was assumed in the original analyses of these data, and the same is done here. However, a serious violation of this assumption could lead to misleading results.

In the advanced colorectal cancer data, the putative surrogate endpoint is PFS (denoted by S) and the true endpoint is OS (denoted by T). PFS is a composite endpoint; it is defined as the minimum of two time-to-event endpoints: $S = \min(\tilde{S}, T)$ where $\tilde{S}$ is the time to progression (TTP). If death occurs before progression ($\tilde{S} > T$ and $S = T$), then $\tilde{S}$ is dependently censored by T. This does not hold in the other direction: death can still occur after progression, but progression cannot occur after death. This data structure has been termed semi-competing risks (Fine et al., 2001) or data following an illness-death model.

In addition to dependent right censoring due to the semi-competing risks nature of the data, there usually is independent right censoring. Let C be the time to *independent* censoring; we thus assume that $C \perp (\tilde{S}, T)|A$. The observed data are then a random sample from $(X, \delta^X, Z, \delta^Z, A)'$ where $X = \min(\tilde{S}, Z)$ is the possibly right-

censored time to progression with event indicator $\delta^X = I(X = \tilde{S})$, and $Z = \min(T, C)$ is the possibly right-censored time to death with event indicator $\delta^Z = I(Z = T)$. Because $\tilde{S}$ is *dependently* censored by the minimum of T and C, and T is only *independently* censored by C, the joint distribution of $(\tilde{S}, T)'$ is only observable on the upper wedge, i.e., the region where $\tilde{S} \leq T$. In addition to being unobservable, the distribution of $(\tilde{S}, T)'$ on the lower wedge is ill defined because progression cannot occur after the patient has died. Therefore, we follow the convention that $\tilde{S} > T$ and $\tilde{S} = \infty$ are equivalent statements where $\tilde{S} = \infty$ indicates that progression never occurs (Nevo and Gorfine, 2022; Xu et al., 2010). This equivalence also implies that $P(\tilde{S} > x, T = x) = P(\tilde{S} = \infty, T = x) = P(S = x, T = x)$.

The joint distribution of PFS and OS, $(S, T)'$, follows from the joint distribution of TTP and OS, $(\tilde{S}, T)'$, because $(S, T)' = \{\min(\tilde{S}, T), T\}'$ (Fu et al., 2013). The D-vine copula model proposed in this work models TTP and OS jointly. This model in turn implies a unique joint distribution for PFS and OS, from which measures of surrogacy will be derived. In other words, $\tilde{S}$ will appear in our modeling strategy, while S will appear in the definition of surrogacy.

3 Information-Theoretic Causal-Inference Framework

3.1 Definition of Surrogacy

In the ITCI framework introduced by Alonso et al. (2015), the surrogate is evaluated in the STS. This framework further assumes that only two treatments are under evaluation in a parallel study design. Consistent with the Neyman-Rubin potential outcomes model, it is presumed that each patient has four potential outcomes represented by $\tilde{\boldsymbol{Y}} = (T_0, \tilde{S}_0, \tilde{S}_1, T_1)'$. Here, $\tilde{S}_a$ and T_a denote the potential outcomes for $\tilde{S}$ and T under treatment $a = 0$ or $a = 1$. The potential outcome for S follows as $S_a = \min(\tilde{S}_a, T_a)$; hence, we can also define the following vector of potential outcomes: $\boldsymbol{Y} = (T_0, S_0, S_1, T_1)'$. The specific ordering of the variables in $\boldsymbol{Y}$ will be clarified in Section 4.2.

The implementation of the Neyman-Rubin model in this paper relies on two fundamental assumptions (Rosenbaum and Rubin, 1983). First, the Stable Unit Treatment Value Assumption (SUTVA)—which encompasses the absence of interference and hidden variations in treatment—establishes the following link between the observed and the potential outcomes:

$$(\tilde{S}, T)' = A \cdot (\tilde{S}_1, T_1)' + (1 - A) \cdot (\tilde{S}_0, T_0)'.$$

Second, the Full Exchangeability Assumption asserts that the potential outcomes are independent of the assigned treatment, that is, $(T_0, \tilde{S}_0, \tilde{S}_1, T_1)' \perp A$. Whereas full exchangeability always holds in randomized trials, SUTVA is an unverifiable assumption that requires justification through subject matter knowledge. Throughout the remainder of this paper, we make both assumptions.

The vector of individual causal treatment effects is defined as $\boldsymbol{\Delta} = (\Delta S, \Delta T)'$, with $\Delta S = S_1 - S_0$ and $\Delta T = T_1 - T_0$. This formulation naturally leads to the following definition of surrogacy in the STS (Alonso, 2018, p. 3).

Definition 1 In the STS, we shall say that S is a good surrogate for T if ΔS conveys a substantial amount of information on ΔT.

The metric that quantifies the extent of "shared" information is referred to as the Individual Causal Association (ICA), which is discussed next (Alonso et al., 2015).

3.2 Individual Causal Association

Mutual information, a well-known metric in information theory, quantifies the degree of shared information between random variables. The mutual information is therefore a well-suited measure to define the ICA, as highlighted by Alonso et al. (2015) and Meyvisch et al. (2020).

Let $P_{\Delta S\Delta T}$, $P_{\Delta S}$, and $P_{\Delta T}$ be the probability measures corresponding to $(\Delta S, \Delta T)'$, ΔS, and ΔT, respectively. Further, let $P_{\Delta S}P_{\Delta T}$ be the product measure of $P_{\Delta S}$ and $P_{\Delta T}$. If $P_{\Delta S\Delta T}$ is absolutely continuous with respect to $P_{\Delta S}P_{\Delta T}$, then the mutual information, $I(\Delta S, \Delta T)$, is defined as the Kullback-Leibler divergence between the corresponding distributions:

$$\begin{aligned} I(\Delta S, \Delta T) &\stackrel{def}{=} D_{KL}\left(P_{\Delta S\Delta T} || P_{\Delta S}P_{\Delta T}\right) \\ &= \int_{\boldsymbol{x}\in \mathrm{R}^2} \log\left(\frac{P_{\Delta S\Delta T}(\boldsymbol{x})}{P_{\Delta S}P_{\Delta T}(\boldsymbol{x})}\right) dP_{\Delta S\Delta T}(\boldsymbol{x}) \end{aligned}$$

where $D_{KL}\left(P_{\Delta S\Delta T} || P_{\Delta S}P_{\Delta T}\right)$ is the Kullback-Leibler divergence and $P_{\Delta S\Delta T}(\boldsymbol{x})/P_{\Delta S}P_{\Delta T}(\boldsymbol{x})$ is the Radon-Nikodym derivative of $P_{\Delta S\Delta T}$ with respect to $P_{\Delta S}P_{\Delta T}$. If the surrogate endpoint is a composite of the true endpoint (e.g., PFS in the advanced colorectal cancer data), it follows that $P(S_a = T_a) > 0$, and similarly that $P(\Delta S = \Delta T) > 0$. As a result, $P_{\Delta S\Delta T}$ includes a probability atom, rendering it not absolutely continuous with respect to $P_{\Delta S}P_{\Delta T}$. Consequently, the mutual information, without any additional modifications, is undefined in this scenario.

To ensure that the mutual information is well-defined when evaluating surrogacy, patients for whom $\Delta S = \Delta T$ can be excluded. The individual causal effects in this subgroup, where $\Delta S \neq \Delta T$, will be denoted by $(\Delta S^*, \Delta T^*)'$. The exclusion of patients exhibiting a perfect association between individual causal treatment effects may seem counterintuitive. However, this perfect association arises directly from S being a composite of T. By excluding these patients in the ICA definition, we quantify the surrogacy of S that is independent of its composite nature. Alternatively, diverging from information theory, one could resort to well-established association measures such as the Spearman correlation.

The mutual information is non-negative, symmetric, invariant under bijective transformations, and equals zero if and only if ΔS^* and ΔT^* are independent. Despite its appealing mathematical properties, the interpretability of the mutual information is hindered due to the absence of an upper bound. This limitation is addressed by mapping $I(\Delta S^*, \Delta T^*)$ into the unit interval, where zero corresponds to independence, and one corresponds to a deterministic relation between ΔS^* and ΔT^*. Ideally, this transformation of the mutual information aligns with well-known measures, such as

the Pearson correlation, under specific distributional assumptions. The squared informational coefficient of correlation (SICC or R_h^2) is a transformation that satisfies the above requirements (Linfoot, 1957; Joe, 1989),

$$R_h^2 = 1 - e^{-2 \cdot I(\Delta S^*, \Delta T^*)}. \tag{1}$$

As previously stated, the ICA can also be defined in terms of common statistical association measures such as the Spearman correlation for $(\Delta S^*, \Delta T^*)'$. The advantage of the SICC over the Spearman correlation, however, is that it is based on the mutual information, a well-known quantity in information theory that is used in many disciplines. Furthermore, metrics based on the mutual information are comparable across settings with diverse scales of measurement (Alonso et al., 2015, 2016; Joe, 1989). Yet, the mutual information—as opposed to the Spearman correlation—is rarely used as a measure of association in applied statistics. Furthermore, the Spearman correlation is always well defined, even in the presence of probability atoms (where the mutual information breaks down). It can therefore also quantify the association for $(\Delta S, \Delta T)'$. Also, the Spearman correlation quantifies monotone association, whereas the SICC quantifies any association. A surrogate for which there is a strong non-monotone association (i.e., large SICC and small Spearman correlation) between the individual causal effects may be of questionable value; the Spearman correlation may then be more appropriate.

There is no general closed-formed formula for computing the ICA regardless of its specific definition. In Section D of the supplementary material, we present a straightforward numerical approach for computing the ICA if the distribution of $\tilde{\boldsymbol{Y}}$ is known. In short, we (i) sample many observations of $(\Delta S^*, \Delta T^*)'$ (or $(\Delta S, \Delta T)'$) by sampling $\tilde{\boldsymbol{Y}}$ and (ii) estimate the ICA in this sample with a consistent estimator. For a sufficiently large number of sampled observations, the resulting estimate approximates the ICA of the known distribution of $\tilde{\boldsymbol{Y}}$.

In the next section, we introduce a D-vine copula model for the distribution of $\tilde{\boldsymbol{Y}}$. From this distribution, the distributions of $(\Delta S, \Delta T)$ and $(\Delta S^*, \Delta T^*)$ follow, from which the ICA follows in turn.

4 Causal-Inference Model: D-Vine Copula Model

In this section, a D-vine copula model for the joint distribution of $\tilde{\boldsymbol{Y}}$ is discussed. First, we briefly introduce (D-vine) copulas. Subsequently, we present the D-vine copula model and the corresponding observed-data likelihood contributions. Next, the partial identifiability of the model is discussed, and finally, methods for evaluating goodness of fit are explored.

4.1 (D-Vine) Copulas

Copulas are functions that allow us to describe the association between random variables independently of their marginal distributions. A bivariate copula is a bivariate distribution function $\mathcal{C}$ with uniform margins: $\mathcal{C} : [0,1]^2 \to [0,1]$. The corresponding

copula density c is obtained by partial differentiation: $c(u,v) = \frac{\partial^2}{\partial u \partial v} \mathcal{C}(u,v)$. In the bivariate setting, numerous parametric copulas are established and commonly applied. However, copulas can be extended to $d > 2$ dimensions. D-vine copulas are a versatile class of d-dimensional copulas that employ bivariate copulas as building blocks (Czado, 2019). The D-vine copula density is defined as the product of a particular set of conditional and unconditional copula densities. A conditional copula density (e.g., $c_{\tilde{S}_0 T_1;\tilde{S}_1}$) is simply the copula density corresponding to a conditional distribution (e.g., $(\tilde{S}_0, T_1)' \,|\, \tilde{S}_1$). A more comprehensive introduction to (D-vine) copulas and related concepts is provided in Section A of the supplementary material.

4.2 D-Vine Copula model

Let $f_{1234} = f_{\tilde{\boldsymbol{Y}}}$ be the joint density function of $\tilde{\boldsymbol{Y}} = (T_0, \tilde{S}_0, \tilde{S}_1, T_1)'$. The copula densities for the bivariate distributions of $(T_0, \tilde{S}_0)'$ and $(\tilde{S}_1, T_1)'$ are denoted by $c_{12} = c_{T_0 \tilde{S}_0}$ and $c_{34} = c_{\tilde{S}_1 T_1}$, respectively. The marginal density functions of T_0, $\tilde{S}_0$, $\tilde{S}_1$, and T_1 are denoted by f_1, f_2, f_3, and f_4, respectively. The D-vine density decomposition of f_{1234} is the product of four marginal densities and six copula densities:

$$f_{1234} = f_1 f_2 f_3 f_4 \cdot c_{12} c_{23} c_{34} \cdot c_{13;2} c_{24;3} \cdot c_{14;23}. \tag{2}$$

The full expressions of the factors in (2) are given in Table 1. The factorization in (2) is a density decomposition without further assumptions. However, without further assumptions, the conditional copula densities in (2) can depend on the conditioning variables in arbitrary ways as long as they correspond to valid copula densities for any fixed value of the conditioning variables, making (2) intractable for constructing models. The simplifying assumption is therefore commonly made in practice and also in this paper (Czado, 2019).

Definition 2 (Simplifying assumption for D-vines) Let Y_i and Y_j correspond to the i'th and j'th variable in $\tilde{\boldsymbol{Y}}$. Let D be a set of indices from $\{1,2,3,4\}$ excluding i and j such that $\boldsymbol{Y_D}$ corresponds to the D'th variable(s) in $\tilde{\boldsymbol{Y}}$. The conditional distribution functions of Y_i and Y_j given $\boldsymbol{Y_D} = \boldsymbol{y_D}$ are denoted by $F_{i|D}$ and $F_{j|D}$, respectively. If

$$c_{i,j;D}\left\{F_{i|D}(y_i|\boldsymbol{y_D}), F_{j|D}(y_j|\boldsymbol{y_D}); \boldsymbol{y_D}\right\} = c_{i,j;D}\left\{F_{i|D}(y_i|\boldsymbol{y_D}), F_{j|D}(y_j|\boldsymbol{y_D})\right\}$$

holds for all $\boldsymbol{y_D}$, and for i, j and D chosen to occur in (2), then the corresponding D-vine distribution is called simplified.

The simplifying assumption for (2) implies that the three conditional copula densities in (2) do not depend on the value of the conditioning variable; this greatly simplifies modeling.

The joint density construction in (2) is very appealing in the ITCI framework for two reasons. First, the association structure is very flexible in a D-vine copula; any copula function can be used for the copulas in (2). Second, because two potential outcomes are systematically missing for each patient, the observed-data likelihood emanates from the bivariate distributions of $(T_0, \tilde{S}_0)'$ and $(\tilde{S}_1, T_1)'$. Conveniently, these

Table 1: Full expressions of the components of the D-vine construction in (2) for $f_{\tilde{\boldsymbol{Y}}}$. The simplifying assumption has been made in the expressions for $c_{T_0,S_1;S_0}$, $c_{S_0,T_1;S_1}$, and $c_{T_0,T_1;S_0,S_1}$.

Component	Expression
f_1	$= f_{T_0}(t_0)$
f_2	$= f_{\tilde{S}_0}(s_0)$
f_3	$= f_{\tilde{S}_1}(s_1)$
f_4	$= f_{T_1}(t_1)$
c_{12}	$= c_{T_0,\tilde{S}_0}\left\{F_{T_0}(t_0), F_{\tilde{S}_0}(s_0)\right\}$
c_{23}	$= c_{\tilde{S}_0,\tilde{S}_1}\left\{F_{\tilde{S}_0}(s_0), F_{\tilde{S}_1}(s_1)\right\}$
c_{34}	$= c_{\tilde{S}_1,T_1}\left\{F_{\tilde{S}_1}(s_1), F_{T_1}(t_1)\right\}$
$c_{13;2}$	$= c_{T_0,\tilde{S}_1;\tilde{S}_0}\left\{F_{T_0\|\tilde{S}_0}(t_0\|\tilde{S}_0 = s_0), F_{\tilde{S}_1\|\tilde{S}_0}(s_1\|\tilde{S}_0 = s_0)\right\}$
$c_{24;3}$	$= c_{\tilde{S}_0,T_1;\tilde{S}_1}\left\{F_{\tilde{S}_0\|\tilde{S}_1}(s_0\|\tilde{S}_1 = s_1), F_{T_1\|\tilde{S}_1}(t_1\|\tilde{S}_1 = s_1)\right\}$
$c_{14;23}$	$= c_{T_0,T_1;\tilde{S}_0,\tilde{S}_1}\left\{F_{T_0\|\tilde{S}_0,\tilde{S}_1}(t_0\|\tilde{S}_0 = s_0, \tilde{S}_1 = s_1), F_{T_1\|\tilde{S}_0,\tilde{S}_1}(t_1\|\tilde{S}_0 = s_0, \tilde{S}_1 = s_1)\right\}$

Table 2: Observable vectors and their associated density functions.

Treatment	Observed Data	Density Function	
$A = 0$	$(T_0, \tilde{S}_0)'$	$f_{T_0,\tilde{S}_0}(t_0, s_0)$	$= c_{T_0,\tilde{S}_0}\left\{F_{T_0}(t_0), F_{\tilde{S}_0}(s_0)\right\} \cdot f_{T_0}(t_0) \cdot f_{\tilde{S}_0}(s_0)$
			$= c_{12}\left\{F_1(t_0), F_2(s_0)\right\} \cdot f_1(t_0) \cdot f_2(s_0)$
$A = 1$	$(\tilde{S}_1, T_1)'$	$f_{\tilde{S}_1,T_1}(s_1, t_1)$	$= c_{\tilde{S}_1,T_1}\left\{F_{\tilde{S}_1}(s_1), F_{T_1}(t_1)\right\} \cdot f_{\tilde{S}_1}(s_1) \cdot f_{T_1}(t_1)$
			$= c_{34}\left\{F_3(s_1), F_4(t_1)\right\} \cdot f_3(s_1) \cdot f_4(t_1)$

bivariate distributions follow immediately from the D-vine construction. The corresponding bivariate density and distribution functions are given, respectively, in Table 2 and Table C.1 in the supplementary material. Note that only the D-vine copula presented here and variations thereof where the copulas for $(T_0, \tilde{S}_0)'$ and $(\tilde{S}_1, T_1)'$ are present in the first tree ($c_{12}c_{23}c_{34}$) lead to analytic expressions for the observed-data likelihood.

4.3 Likelihood Contributions

In the remainder of this paper, we focus on parametric estimation of the D-vine copula model. The generic likelihood contributions for the observed data are given next, taking into account that only $(T_0, \tilde{S}_0)'$ or $(\tilde{S}_1, T_1)'$ are observed with possible right-censoring. We further use survival copulas (denoted by a tilde), which simplifies the expressions for right-censored data. Let $\tilde{\mathcal{C}}$ be a copula, a bivariate survival function can then be defined as follows (Georges et al., 2001):

$$P(X > x, Y > y) = \tilde{\mathcal{C}}\left\{S_X(x), S_Y(y)\right\},$$

where S_X and S_Y are marginal survival functions. The survival copula, $\tilde{\mathcal{C}}$, is further related to the copula $\mathcal{C}$ of the distribution of $(X,Y)'$ as follows:

$$\tilde{\mathcal{C}}(u,v) = u + v - 1 + \mathcal{C}(1-u, 1-v).$$

Following the notation introduced in Section 2, let the observed vector for patient i be $(x_i, \delta_i^x, z_i, \delta_i^z, a_i)$. Assuming SUTVA, full exchangeability, and independent censoring, the possible likelihood contributions for patient i with $a_i = 1$ are as follows:

- Both the surrogate and the true endpoint event are observed, $\delta_i^x = \delta_i^z = 1$:

$$L_i = f_{\tilde{S}_1}(x_i) \cdot f_{T_1}(z_i) \cdot \tilde{c}_{\tilde{S}_1, T_1}\left\{S_{\tilde{S}_1}(x_i), S_{T_1}(z_i)\right\}$$

- The surrogate event is observed and the true endpoint is right-censored, $\delta_i^x = 1$ and $\delta_i^z = 0$:

$$L_i = f_{\tilde{S}_1}(x_i) \cdot \frac{\partial \tilde{\mathcal{C}}_{\tilde{S}_1, T_1}\left\{u, S_{T_1}(z_i)\right\}}{\partial u}\Big|_{u = S_{\tilde{S}_1}(x_i)}$$

- Both endpoints are right-censored, $\delta_i^x = \delta_i^z = 0$:

$$L_i = \tilde{\mathcal{C}}_{\tilde{S}_1, T_1}\left\{S_{\tilde{S}_1}(x_i), S_{T_1}(z_i)\right\}$$

- The true endpoint event occurs before the surrogate event, $\delta_i^x = 0$ and $\delta_i^z = 1$:

$$L_i = f_{T_1}(z_i) \cdot \frac{\partial \tilde{\mathcal{C}}_{\tilde{S}_1, T_1}\left\{S_{\tilde{S}_1}(x_i), v\right\}}{\partial v}\Big|_{v = S_{T_1}(z_i)}.$$

For the marginal densities and survival functions, any (survival) distribution can be used. Similarly, for the survival copulas, any copula can be used. Note that by definition of the variables in the previous subsection, $x_i > z_i$ cannot occur. The derivations of the above expressions are given in Section B.2 of the supplementary material.

Table 3: Interpretation of the unidentifiable components of the D-vine density f_{1234} in (2). The full expressions for the individual components are given in Table B.1 in the supplementary material.

Component	Interpretation
c_{23}	Dependence structure of $(\tilde{S}_0, \tilde{S}_1)$
$c_{13;2}$	Dependence structure of $(T_0, \tilde{S}_1)\|\tilde{S}_0$
$c_{24;3}$	Dependence structure of $(\tilde{S}_0, T_1)\|\tilde{S}_1$
$c_{14;23}$	Dependence structure of $(T_0, T_1)\|\tilde{S}_0, \tilde{S}_1$

4.4 Identifiability

The copulas c_{23}, $c_{13;2}$, $c_{24;3}$, and $c_{14;23}$ do not appear in the observed-data likelihood contributions and thus cannot be estimated from the observed data: they are unidentifiable. The interpretation of the unidentifiable components in (2) is summarized in Table 3. The partial identifiability of the model is addressed by a formal sensitivity analysis, as explained in Section 5.

4.5 Goodness of fit

We will estimate the D-vine copula model parametrically in Section 6; therefore, it is important to evaluate the goodness of fit. Although this model is only partially identifiable, one can still evaluate the goodness of fit for the observable bivariate margins. This is split into two parts: (i) goodness of fit for the marginal survival functions and (ii) goodness of fit for the identifiable bivariate copulas, i.e., $\mathcal{C}_{12}$ and $\mathcal{C}_{34}$. First, to evaluate the appropriateness of the marginal survival functions, $P(S_a > t)$ and $P(T_a > t)$, the model-based estimates are compared with the corresponding Kaplan-Meier estimates. Second, to evaluate the appropriateness of the identifiable bivariate copulas, we focus on the conditional distribution of $S_a|T_a$. Specifically, we compare the model-based estimates with semi-parametric estimates of $E(S_a|T_a = t, S_a < T_a)$ and $P(S_a = t|T_a = t)$. Further, $E(S_a|T_a = t, S_a < T_a)$ is estimated semi-parametrically by regressing X on Z in the subset of patients with $\delta^X = \delta^Z = 1$ and $A = a$. Similarly, $P(S_a = t|T_a = t)$ is estimated semi-parametrically by regressing $I(X = Z)$ on Z in the subset of patients with $\delta^Z = 1$ and $A = a$. In Section E of the supplementary material, further details are provided on how to derive these functions from the estimated D-vine copula model and on how to estimate these quantities semi-parametrically.

5 Sensitivity Analysis

Estimating the ICA is challenging because the underlying model is only partially identifiable. This challenge can be tackled by incorporating a sensitivity analysis into the estimation process. It then becomes crucial to do the sensitivity analysis within a formal framework that enables the application of classical inferential tools or their extensions.

In the remainder of this section, we formalize the partial identifiability of the ICA using three concepts that are fundamental in a formal sensitivity analysis: *imprecision*, *ignorance*, and *uncertainty*. These concepts were introduced by Molenberghs et al. (2001) and Kenward et al. (2001), and studied in detail by Vansteelandt et al. (2006). To keep the paper concise, we only repeat the details from Vansteelandt et al. (2006) that are necessary to understand the current methods. For more theoretical details, we refer the interested reader to Vansteelandt et al. (2006).

5.1 Partial Identifiability

In the previous section, we assumed a D-vine copula model for $\tilde{\boldsymbol{Y}} \sim F_{\tilde{\boldsymbol{Y}}}$ indexed by two variationally independent parameters: $\boldsymbol{\beta} = (f_1, f_2, f_3, f_4, \mathcal{C}_{12}, \mathcal{C}_{34})'$ and $\boldsymbol{\nu} = (\mathcal{C}_{23}, \mathcal{C}_{13;2}, \mathcal{C}_{24;3}, \mathcal{C}_{14;23})'$ where $\boldsymbol{\beta}$ is identifiable and $\boldsymbol{\nu}$ is unidentifiable. The true value for the identifiable parameter is further denoted by $\boldsymbol{\beta_0}$. The unidentifiability of $\boldsymbol{\nu}$ arises from a structural missing data problem: due to the fundamental problem of causal inference, only 50% of the elements of $\tilde{\boldsymbol{Y}}$ are observable. Notice further that the ICA is a functional of $F_{\Delta S \Delta T}$ (or $F_{\Delta S^* \Delta T^*}$), which is itself a functional of $F_{\tilde{\boldsymbol{Y}}}$ indexed by $(\boldsymbol{\beta}', \boldsymbol{\nu}')'$. To highlight the partial identifiability of the ICA, one can express it as a function of the specified parameters: $ICA = R_h^2(\boldsymbol{\beta}, \boldsymbol{\nu})$.

5.2 Intervals of Ignorance and Uncertainty

5.2.1 Imprecision, Ignorance and Uncertainty

We distinguish between *imprecision* due to sampling variability on the one hand, and *ignorance* due to unidentifiability on the other hand.

Let $\hat{\boldsymbol{\beta}}_N$ be a consistent estimator for $\boldsymbol{\beta_0}$ indexed by the sample size N and assume further that the true value for the sensitivity parameter, denoted by $\boldsymbol{\nu_0}$, is known. Then it follows from the continuous mapping theorem that $R_h^2(\hat{\boldsymbol{\beta}}_N, \boldsymbol{\nu_0})$ is a consistent estimator if $R_h^2(\boldsymbol{\beta}, \boldsymbol{\nu_0})$ is a continuous function of $\boldsymbol{\beta}$. In finite samples, the variability in $R_h^2(\hat{\boldsymbol{\beta}}_N, \boldsymbol{\nu_0})$ is a consequence of the sampling variability in $\hat{\boldsymbol{\beta}}_N$; this variability is termed *imprecision* and it can be quantified using standard errors and confidence intervals.

The observed data, however, convey no information about the sensitivity parameter, leaving all values for $\boldsymbol{\nu}$ that yield a valid distribution within the realm of possibility. Nonetheless, one may constrain this parameter to a plausible subset, denoted by Γ. Note that, even if the sample size N tends to infinity, and precise knowledge of $\boldsymbol{\beta_0}$ is thus available, the exact value of the ICA remains unknown. This type of uncertainty emanating from the unidentifiability of $\boldsymbol{\nu}$ is called *ignorance*. To address this problem, one can define "all values for the ICA compatible with the observable data and any additional assumptions", using the interval

$$\begin{aligned} ir(ICA, \Gamma) &= [ICA_l, ICA_u], \\ ICA_l &= \min\left\{R_h^2(\boldsymbol{\beta_0}, \boldsymbol{\nu}) : \boldsymbol{\nu} \in \Gamma\right\}, \\ ICA_u &= \max\left\{R_h^2(\boldsymbol{\beta_0}, \boldsymbol{\nu}) : \boldsymbol{\nu} \in \Gamma\right\}, \end{aligned}$$

which is termed the *interval of ignorance*. The bounds of this interval are sharp in the sense that (i) $ICA_u \geq R_h^2(\boldsymbol{\beta_0}, \boldsymbol{\nu}) \forall \boldsymbol{\nu} \in \Gamma$ and (ii) $\exists \boldsymbol{\nu}_u \in \Gamma \; s.t. \; ICA_u = R_h^2(\boldsymbol{\beta_0}, \boldsymbol{\nu}_u)$ (and similarly for ICA_l.)

In practice, an infinitely large sample is not available, and the true value of the sensitivity parameter remains unknown. Therefore, inferences about the ICA must integrate both imprecision and ignorance, a combination often termed *uncertainty* (Molenberghs and Kenward, 2007, Ch. 21).

5.2.2 Estimated Interval of Ignorance

In the definition of the interval of ignorance, the true identifiable parameter is replaced with its estimate, $\hat{\boldsymbol{\beta}}_N$, to obtain the *estimated interval of ignorance*:

$$\begin{aligned} \widehat{ir}_N(ICA, \Gamma) &= [\widehat{ICA}_l, \widehat{ICA}_u], \\ \widehat{ICA}_l &= \min \left\{ R_h^2(\hat{\boldsymbol{\beta}}_N, \boldsymbol{\nu}) : \boldsymbol{\nu} \in \Gamma \right\}, \\ \widehat{ICA}_u &= \max \left\{ R_h^2(\hat{\boldsymbol{\beta}}_N, \boldsymbol{\nu}) : \boldsymbol{\nu} \in \Gamma \right\}. \end{aligned} \tag{3}$$

This interval is an estimate for the true ICA, denoted by R_{h0}^2. Since we are estimating a scalar parameter using an interval, the definition of (weak) consistency of point estimators should be extended to interval estimators. Intuitively, $R_h^2(\hat{\boldsymbol{\beta}}_N, \boldsymbol{\nu})$ should be consistent for $R_h^2(\boldsymbol{\beta_0}, \boldsymbol{\nu})$ for all $\boldsymbol{\nu} \in \Gamma$. Vansteelandt et al. (2006, Definition 2) formalize this notion of weak convergence. Their definition is adapted to the present problem with some modifications.

Definition 3 (Weak consistency of $\hat{ir}_N$) The estimated interval of ignorance, as defined in (3), is termed weakly consistent for R_{h0}^2 if

$$R_h^2(\hat{\boldsymbol{\beta}}_N, \boldsymbol{\nu}) \xrightarrow{P} R_h^2(\boldsymbol{\beta_0}, \boldsymbol{\nu}) \; \forall \; \boldsymbol{\nu} \in \Gamma,$$

where $\xrightarrow{P}$ denotes convergence in probability.

If $R_h^2(\boldsymbol{\beta}, \boldsymbol{\nu})$ is a continuous function of $\boldsymbol{\beta}$ (treating $\boldsymbol{\nu}$ as fixed), consistency of $\hat{\boldsymbol{\beta}}_N$ implies consistency of $R_h^2(\hat{\boldsymbol{\beta}}_N, \boldsymbol{\nu})$ by the continuous mapping theorem. Hence, the estimated interval of ignorance will generally be weakly consistent for R_{h0}^2. Clearly, this is only meaningful when $\boldsymbol{\nu_0} \in \Gamma$, that is, when the true sensitivity parameter lies within the plausible set.

Weak consistency of the estimated interval of ignorance is a very useful property. Indeed, it implies that this estimated interval will cover the truth with arbitrarily large probability as the sample size increases if $\boldsymbol{\nu_0} \in \Gamma$. However, this does not provide any guarantee in finite samples; this is addressed by the pointwise uncertainty interval.

5.2.3 Pointwise Uncertainty Interval

In the presence of unidentifiable parameters, the classical confidence interval can be extended to the *pointwise uncertainty interval* (UI_p). This interval covers the

$R_h^2(\boldsymbol{\beta_0}, \boldsymbol{\nu})$ with at least $1-\alpha$ probability for all $\boldsymbol{\nu} \in \Gamma$. Therefore, if Γ contains the true sensitivity parameter, the UI_p will cover the $R_h^2(\boldsymbol{\beta_0}, \boldsymbol{\nu_0})$ with at least $1-\alpha$ probability. This suggests that UI_p can be interpreted as a traditional confidence interval, provided that $\boldsymbol{\nu_0}$ lies in Γ. As a result, this interval can be used to test the null hypothesis $H_0 : R_h^2 = R_{h0}^2$ against the alternative $H_1 : R_h^2 \neq R_{h0}^2$ at the significance level α. As for identifiable parameters, we reject H_0 when R_{h0}^2 falls outside the UI_p.

Definition 4 (Pointwise Uncertainty Interval, UI_p) $UI_p(ICA, \Gamma)$ is a $1-\alpha$ pointwise uncertainty interval for R_{h0}^2 when its pointwise coverage probability is at least $1-\alpha$, i.e.,

$$P_{\boldsymbol{\beta_0}}\left\{R_h^2(\boldsymbol{\beta_0}, \boldsymbol{\nu}) \in UI_p(ICA, \Gamma)\right\} \geq 1-\alpha, \; \forall \, \boldsymbol{\nu} \in \Gamma,$$

where $P_{\boldsymbol{\beta_0}}$ means that the probability is with respect to the true observable distribution.

To construct pointwise uncertainty intervals, the following assumption is needed (Vansteelandt et al., 2006, Assumption 2).

Assumption 1 The sensitivity parameters, $\boldsymbol{\nu}_l$ and $\boldsymbol{\nu}_u$, in Γ that correspond to the lower bound $ICA_l = R_h^2(\boldsymbol{\beta}, \boldsymbol{\nu}_l)$ and the upper bound $ICA_u = R_h^2(\boldsymbol{\beta}, \boldsymbol{\nu}_u)$ are independent of $\boldsymbol{\beta}$.

However, this assumption may frequently be violated. If this assumption fails, Vansteelandt et al. (2006) suggested that one could still determine $\boldsymbol{\nu}_l$ and $\boldsymbol{\nu}_u$ in an adaptive way by setting $\boldsymbol{\beta_0}$ to its estimated value and then proceeding as if the adaptive estimates for $\boldsymbol{\nu}_l$ and $\boldsymbol{\nu}_u$, denoted by $\hat{\boldsymbol{\nu}}_l$ and $\hat{\boldsymbol{\nu}}_u$, are fixed. In the current setting, no general closed-form expressions are available for $\hat{\boldsymbol{\nu}}_l$ and $\hat{\boldsymbol{\nu}}_u$. These quantities are approximated using the Monte Carlo approach discussed in the next section.

Following the approach outlined in Vansteelandt et al. (2006), it can be demonstrated that $[C_l, C_u]$ constitutes a $1-\alpha$ pointwise uncertainty interval for the ICA, where C_l (and C_u) represents the lower (and upper) limit of a $1-\alpha^*/2$ one-sided confidence interval for $R_h^2(\boldsymbol{\beta_0}, \boldsymbol{\nu}_l)$ (and $R_h^2(\boldsymbol{\beta_0}, \boldsymbol{\nu}_u)$). Here, α^* corresponds to the solution of Equation (4.3) in Vansteelandt et al. (2006). However, these authors also showed that $\alpha^*/2 \approx \alpha$ if the level of ignorance is substantial compared to the imprecision. As long as the sample size is sufficiently large, this holds in the ITCI framework because at least 50% of the potential outcomes are missing.

5.3 Implementation of the Sensitivity Analysis

From a mathematical standpoint, the ICA can be viewed as a function of a vector of unidentifiable parameters, denoted by $R_h^2(\boldsymbol{\beta_0}, \boldsymbol{\nu}) : \Gamma \rightarrow [0,1]$. This mathematical function fully encapsulates the surrogate's validity across all scenarios compatible with the observable data. Theoretically, one could explore the behavior of $R_h^2(\boldsymbol{\beta_0}, \boldsymbol{\nu})$ in a purely mathematical manner, but the dimensionality of $\boldsymbol{\nu}$ and its complex, non-linear relationship with the ICA render this problem mathematically intractable. A more practical approach involves studying this functional relationship using a stochastic procedure, where a sufficiently large number of $\boldsymbol{\nu}$ vectors in Γ are

sampled. For each element in this sample, the joint distribution of $\tilde{\boldsymbol{Y}}$ can be determined, followed by the joint distribution of the individual causal treatment effects $\boldsymbol{\Delta} = (\Delta S, \Delta T)'$ (or $\boldsymbol{\Delta}^* = (\Delta S^*, \Delta T^*)'$) and the corresponding $R^2_h(\boldsymbol{\beta_0}, \boldsymbol{\nu})$. The frequency distribution of the resulting $R^2_h(\boldsymbol{\beta_0}, \boldsymbol{\nu})$ values describes the behavior of the ICA on Γ and serves as a sensitivity analysis to evaluate the validity of the putative surrogate across numerous plausible $\boldsymbol{\nu}$ vectors. The previously introduced interval of ignorance defines the range of $R^2_h(\boldsymbol{\beta_0}, \boldsymbol{\nu})$ values within Γ, offering a succinct summary of the function's behavior across the plausible set. This approach can be implemented using the following algorithm once $\boldsymbol{\beta_0}$ is fixed to its estimated value:

1. Given $F_{\boldsymbol{\nu},\Gamma}$, a distribution for $\boldsymbol{\nu}$ with support Γ, independently sample $\boldsymbol{\nu}^{(l)}$ L times. Refer to Section C.2 of the supplementary material for the proposed sampling method.
2. Compute the ICA for each sampled sensitivity parameter while holding the identifiable parameter fixed at its estimated value: $R^2_h(\hat{\boldsymbol{\beta}}_N, \boldsymbol{\nu}^{(l)})$ for $l = 1, ..., L$.
3. Properly summarize the set $\{R^2_h(\hat{\boldsymbol{\beta}}_N, \boldsymbol{\nu}^{(l)}) : l = 1, ..., L\}$

In this procedure, we determine $\hat{\boldsymbol{\nu}}_l$ and $\hat{\boldsymbol{\nu}}_u$ as the sensitivity parameters corresponding to the minimum and maximum values obtained for the ICA. The corresponding approximated estimated interval of ignorance summarizes the frequency distribution of $R^2_h(\hat{\boldsymbol{\beta}}_N, \hat{\nu}^{(l)})$ as $[R^2_h(\hat{\boldsymbol{\beta}}_N, \hat{\boldsymbol{\nu}}_l), R^2_h(\hat{\boldsymbol{\beta}}_N, \hat{\boldsymbol{\nu}}_u)]$. This is only an approximation if L is finite; hence, L should be sufficiently large for this approximation to be accurate. Furthermore, imprecision can be incorporated into the analysis by computing the pointwise uncertainty interval. This involves calculating the one-sided $1-\alpha$ confidence intervals for $R^2_h(\boldsymbol{\beta_0}, \hat{\nu}_l)$ and $R^2_h(\boldsymbol{\beta_0}, \hat{\nu}_u)$ using a parametric bootstrap, as explained next.

In the parametric bootstrap, the starting assumption is that the estimator for $\boldsymbol{\beta_0}$ is consistent and asymptotically normal:

$$N^{1/2}(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta_0}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}),$$

where $\xrightarrow{d}$ denotes convergence in distribution. Subsequently, a bootstrap procedure is conducted by resampling the estimated identifiable parameter B times from the *estimated* sampling distribution:

$$\hat{\boldsymbol{\beta}}^{(b)} \sim N\left(\hat{\boldsymbol{\beta}}_N, \frac{\hat{\boldsymbol{\Sigma}}_N}{N}\right),$$

where $\hat{\boldsymbol{\Sigma}}_N$ is a consistent estimator for $\boldsymbol{\Sigma}$. Further, bootstrap replicates for the ICA are generated with the sensitivity parameter set at $\hat{\boldsymbol{\nu}}_l$ as $R^2_h(\hat{\boldsymbol{\beta}}^{(b)}, \hat{\boldsymbol{\nu}}_l)$ (and similarly for $\hat{\boldsymbol{\nu}}_u$). Finally, let $F_{R^2_h}(\cdot; \hat{\boldsymbol{\nu}}_l)$ and $F_{R^2_h}(\cdot; \hat{\boldsymbol{\nu}}_u)$ be the distribution functions of $R^2_h(\hat{\boldsymbol{\beta}}^{(b)}, \hat{\boldsymbol{\nu}}_l)$ and $R^2_h(\hat{\boldsymbol{\beta}}^{(b)}, \hat{\boldsymbol{\nu}}_u)$, respectively. Then the limits of the $1-\alpha$ one-sided confidence intervals for $R^2_h(\boldsymbol{\beta_0}, \hat{\boldsymbol{\nu}}_l)$ and $R^2_h(\boldsymbol{\beta_0}, \hat{\boldsymbol{\nu}}_u)$—these correspond to C_l and C_u in the pointwise uncertainty interval—are $F^{-1}_{R^2_h}(\alpha; \hat{\boldsymbol{\nu}}_l)$ and $F^{-1}_{R^2_h}(1-\alpha; \hat{\boldsymbol{\nu}}_u)$, respectively.

5.4 Bayesian Interpretation

The main goal of the sensitivity analysis is to compute the frequency distribution that describes the behavior of the ICA on Γ, along with the estimated interval of ignorance and the pointwise uncertainty interval that summarize it. However, this frequency distribution also lends itself to a Bayesian interpretation. Specifically, if we define $F_{\boldsymbol{\nu},\Gamma}$ as a prior distribution for the unidentifiable parameters, then the set of values for $R_h^2(\hat{\boldsymbol{\beta}}_N, \boldsymbol{\nu}^{(l)})$ obtained from the sensitivity analysis represents a draw from the empirical posterior distribution of the ICA, disregarding the sampling variability in $\hat{\boldsymbol{\beta}}_N$. Alternatively, a fully Bayesian approach, as proposed by Conlon et al. (2017), could also be implemented.

6 Application: Advanced Colorectal Cancer

The pooled data on advanced colorectal cancer, described in Section 2, were used by Burzykowski et al. (2001) to evaluate the surrogacy of PFS for OS within the meta-analytic framework, with center serving as the unit of analysis. Their findings indicated that PFS does not serve as a valid surrogate for OS at either the trial or the individual level in advanced colorectal cancer. We re-evaluate these data using the methods outlined in this paper. For brevity, only a summary of the main findings is provided here, with no mention of software. A comprehensive analysis, along with references to the R package *Surrogate*, is available in Section F of the supplementary material.

6.1 Estimated Model

As outlined in Section 4.2, the D-vine copula model is characterized by two components: (i) the copula densities and (ii) the marginal distributions. In total, four parametric copulas are explored: Gaussian, Frank, Gumbel, and Clayton (Nelsen, 2006). The same type of copula is assumed for all copula densities in (2); however, each copula density possesses its own unique parameter. Furthermore, the Royston-Parmar model, devoid of covariates, is employed to model the marginal distributions (Royston and Parmar, 2002). The flexibility of this model is modulated by adjusting the number of internal knots. A distinct Royston-Parmar model is employed for each of the four marginal survival functions with a common number of internal knots.

The D-vine copula model is estimated with maximum likelihood (using the BFGS algorithm implemented by `optim()` in R) for all combinations of the four copula types and two to five internal knots. Among the fitted models, the one with Gaussian copulas and two internal knots is chosen due to its lowest corresponding AIC value. The goodness-of-fit plots for the control and experimental group are depicted in Fig. 1 and 2, respectively. The selected model adequately captures the main features of the control group data, although there seems to be a slight lack of fit for the experimental group. A one-parameter copula may not be sufficiently flexible to fully capture the association between PFS and OS in these data. Further details on all fitted

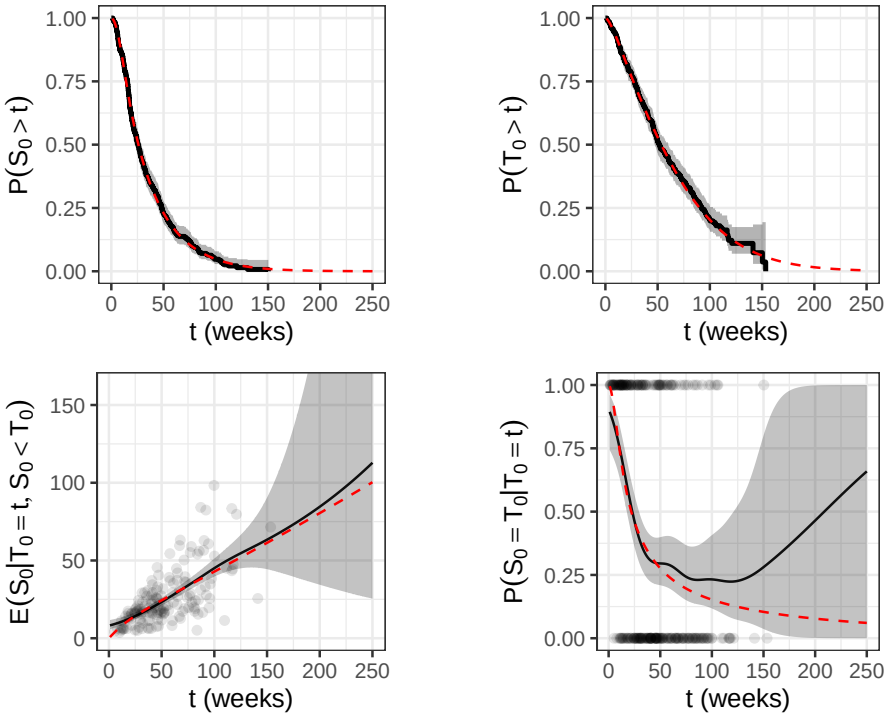

Fig. 1: Goodness-of-fit plots comparing the model-based and semi/non-parametric estimates of $P(S_0 > t)$, $P(T_0 > t)$, $E(S_0|T_0 = t, S_0 < T_0)$, and $P(S_0 = T_0|T_0 = t)$ for the control group. The black lines are the Kaplan-Meier estimates (upper plots) and the semi-parametric estimates (lower plots); the shaded regions are corresponding 95% pointwise confidence intervals. The semi-parametric models are additive models fitted using `mgcv::gam(z~s(x), family = quasi("log", "mu")` for the lower left plot and `mgcv::gam(z~s(x), family = binomial())` for the lower right plot. The red dashed lines are the model-based estimates.

models, along with accompanying goodness-of-fit plots, are provided in Section F of the supplementary material.

6.2 Assumptions

As previously mentioned, additional assumptions can be made about the sensitivity parameter, resulting in a smaller plausible set, Γ. While such constraints are predicated on unverifiable assumptions, it is often feasible to rule out certain scenarios based on domain knowledge. Subsequently, all unverifiable assumptions in the analysis of the advanced colorectal cancer data are enumerated.

The first unverifiable assumption pertains to the parametric form of the unidentifiable components in (2), namely c_{23}, $c_{13;2}$, $c_{24;3}$, and $c_{14;23}$.

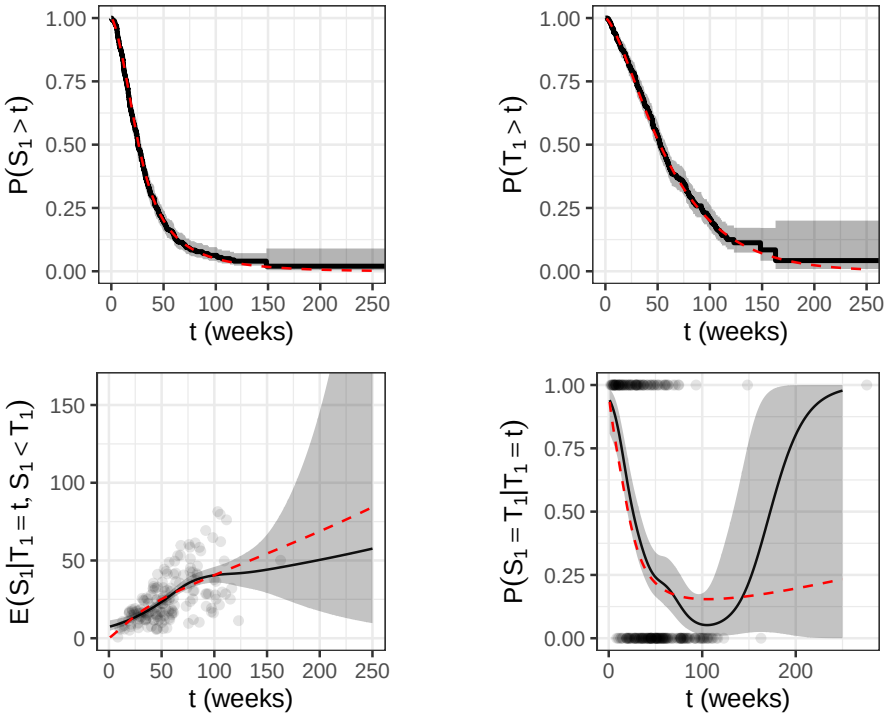

Fig. 2: Goodness-of-fit plots comparing the model-based and semi/non-parametric estimates of $P(S_1 > t)$, $P(T_1 > t)$, $E(S_1|T_1 = t, S_1 < T_1)$, and $P(S_1 = T_1|T_1 = t)$ for the experimental group. The elements of these plots are described in the caption of Fig. 1.

Assumption 2 c_{23}, $c_{13;2}$, $c_{24;3}$, and $c_{14;23}$ are Gaussian copulas satisfying the simplifying assumption.

In addition to imposing modeling assumptions on the unidentifiable copulas, we constrain the parameter space for these copulas based on domain knowledge. The most important factor determining TTP in advanced colorectal cancer is the patient's disease severity. This can be conceptualized by a shared frailty, which produces a positive correlation between $\tilde{S}_0$ and $\tilde{S}_1$. The magnitude of this positive correlation is related to the variance of the frailty term and is anticipated to be quite strong. Consequently, in the sensitivity analysis, we may eliminate scenarios with negative or weakly positive associations but cannot disregard very strong positive associations. Therefore, we stipulate the following assumption regarding the Spearman correlation between $\tilde{S}_0$ and $\tilde{S}_1$, denoted by $\rho_{\tilde{S}_0,\tilde{S}_1}$.

Assumption 3 $\rho_{\tilde{S}_0,\tilde{S}_1} \in [0.5, 0.95]$.

Conditional independence assumptions are also invoked in the analysis. These assumptions, which have been employed in other frameworks for evaluating surrogate endpoints (e.g., by Parast et al. (2016)), are as follows:

Assumption 4 (Conditional Independence) $\tilde{S}_0 \perp T_1 \,|\, \tilde{S}_1$ and $\tilde{S}_1 \perp T_0 \,|\, \tilde{S}_0$.

The first conditional independence statement can be interpreted as: "The TTP under the control treatment does not yield information about the survival time under the experimental treatment if we already know the TTP under the experimental treatment." The second conditional independence assumption can be interpreted in a similar way.

Lastly, we introduce additional assumptions about the association in $(T_0, T_1)' \,|\, \tilde{S}_0, \tilde{S}_1$. This association describes the relation between the potential survival times given TTP under both treatments. It is thus (indirectly) conditional on unobservable factors influencing disease severity and treatment response. Consequently, by indirectly controlling for the most influential factors inducing a strong positive association between T_0 and T_1, we invoke another type of frailty argument to posit that this association is likely to be (weakly) positive. This frailty encapsulates patient characteristics unrelated to cancer directly, such as cardiovascular health. Indeed, cardiovascular health is not (directly) linked to cancer or $(\tilde{S}_0, \tilde{S}_1)'$, but it does impact survival. Since we indirectly condition on cancer-related causes of death, the conditional association is expected to be weaker than the unconditional association. These assumptions can be seamlessly integrated into the sensitivity analysis since this conditional association is modeled by $c_{14;23}$ in (2), which is determined by the corresponding Spearman correlation $\rho_{T_0,T_1|\tilde{S}_0,\tilde{S}_1}$.

Assumption 5 $\rho_{T_0,T_1|\tilde{S}_0,\tilde{S}_1} \in [0.2, 0.8]$

In the ensuing sensitivity analysis, we incorporate the aforementioned four unverifiable assumptions.

6.3 Results

Under the aforementioned assumptions, the estimated interval of ignorance for the ICA, as quantified by the SICC between ΔS^* and ΔT^*, spans from 0.201 to 0.480 based on $L = 5000$ replications in the sensitivity analysis. Correspondingly, the 95% pointwise uncertainty interval is from 0.175 to 0.549. This suggests, according to Definition 1, that PFS serves only as a moderate surrogate for OS in advanced colorectal cancer. The sensitivity analysis and the parametric bootstrap took about 50 minutes using 10 (virtual) cores on an Intel Core i7 10750H processor.

Assumption 2 is a parametric assumption about the unverifiable parts of the association structure of $\tilde{\boldsymbol{Y}}$. Other parametric copulas could have been used in this assumption; furthermore, it is difficult to justify any choice between different unidentifiable parametric copulas. We therefore repeated the sensitivity analysis where the Gaussian copula in Assumption 2 is replaced with the Frank, Gumbel, or Clayton copula. These results are summarized in Table 4, which reveals that the intervals are slightly sensitive to the parametric copula in Assumption 2. The sensitivity analysis was also repeated with relaxed assumptions 3–5 (described in Section F.5 of the supplementary

material). As expected, this adjustment can substantially influence the results, though it has a relatively minor impact on the upper bound of the uncertainty interval. In Section F.4.2 of the supplementary material, we dive deeper into these assumptions by examining their implications for (relations between) easy-to-interpret quantities.

Table 4: Results for the sensitivity analysis under different parametric choices in Assumption 2. Note that these results are based on the same model for the observable data, described in Section 6.1, with Gaussian identifiable copulas ($\mathcal{C}_{T_0,\tilde{S}_0}$ and $\mathcal{C}_{\tilde{S}_1,T_1}$) and separate Royston-Parmar models with two knots for the marginal distributions of $\tilde{S}_a$ and T_a. Est. II: Estimated interval of ignorance; IU: Uncertainty interval.

Unid. Copulas	$R^2_h(\Delta S^*, \Delta T^*)$		$\rho_{sp}(\Delta S^*, \Delta T^*)$		$\rho_{sp}(\Delta S, \Delta T)$	
	Est. II	95% IU	Est. II	95% IU	Est. II	95% IU
Gaussian	(0.20, 0.48)	(0.18, 0.55)	(0.37, 0.65)	(0.30, 0.70)	(0.50, 0.71)	(0.45, 0.74)
Frank	(0.22, 0.49)	(0.20, 0.55)	(0.38, 0.64)	(0.32, 0.69)	(0.51, 0.70)	(0.47, 0.73)
Gumbel	(0.26, 0.54)	(0.23, 0.60)	(0.44, 0.71)	(0.33, 0.75)	(0.56, 0.76)	(0.49, 0.79)
Clayton	(0.17, 0.45)	(0.15, 0.52)	(0.28, 0.58)	(0.23, 0.63)	(0.41, 0.63)	(0.37, 0.68)

When quantifying the ICA using the Spearman correlation, one can explore the association between ΔS^* and ΔT^*, or the association between ΔS and ΔT. The corresponding estimated intervals of ignorance are from 0.368 to 0.654 and from 0.497 to 0.707, respectively, with 95% pointwise uncertainty intervals from 0.300 to 0.700 and from 0.446 to 0.743. Hence, the measure of surrogacy only decreases slightly when we account for the composite nature of PFS (by using $\rho_{sp}(\Delta S^*, \Delta T^*)$ instead of $\rho_{sp}(\Delta S, \Delta T)$).

The ICA values (quantified by the SICC between ΔS^* and ΔT^*) from the sensitivity analysis are summarized in Fig. 3. The frequency distribution indicates that the ICA tends to take relatively small values on Γ, suggesting that PFS has limited value as a surrogate for OS in this scenario. As noted earlier, this empirical distribution can also be interpreted as an empirical posterior distribution for the ICA when sampling variability in $\hat{\boldsymbol{\beta}}_N$ is ignored. The choice between a Bayesian or frequentist interpretation has little influence on the conclusion in this example.

7 Discussion and Limitations

7.1 Discussion

In this paper, the ITCI framework for evaluating surrogate endpoints has been extended to time-to-event endpoints using the flexibility of D-vine copulas. Because the ICA is not identifiable from the observed data, we proposed a sensitivity analysis whose results are summarized by the estimated interval of ignorance and the $1-\alpha$ pointwise uncertainty interval. These methods were applied to clinical trial data in advanced colorectal cancer; the results exclude PFS as a strong surrogate for OS in this area.

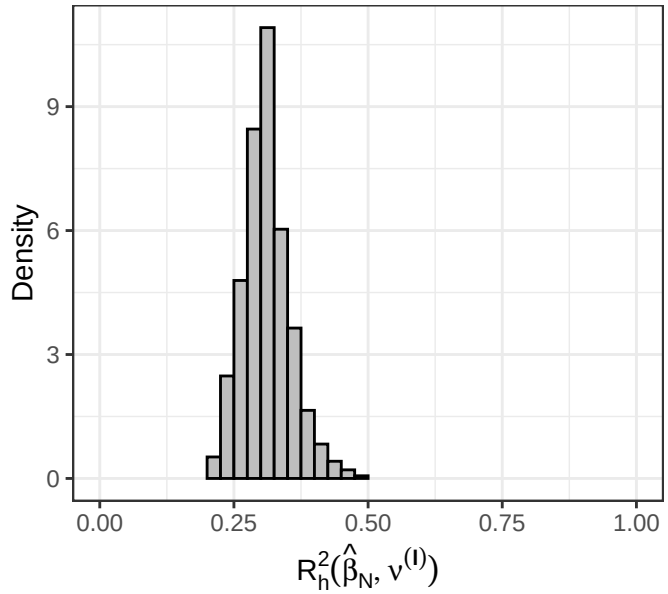

Fig. 3: Histogram of the $R_h^2(\hat{\boldsymbol{\beta}}_N, \boldsymbol{\nu}^{(l)})$ values obtained from the sensitivity analysis. As detailed in the text, this histogram can be interpreted in a Bayesian fashion if we let $F_{\boldsymbol{\nu},\Gamma}$ be a prior distribution and ignore sampling variability. R_h^2: squared informational coefficient of correlation.

It is difficult to provide heuristics regarding the magnitude of the ICA that is sufficient for a surrogate to be deemed valid. Nonetheless, we can borrow heuristics for the R^2_{trial} in the meta-analytic framework for two reasons (Alonso et al., 2017). First, the ICA is quantified by the SICC which is equal to the squared Pearson correlation if $(\Delta S, \Delta T)'$ is bivariate normal. Second, the ICA is the patient-level analog of the R^2_{trial}. Indeed, the ICA quantifies the association between *patient-level* treatment effects while the R^2_{trial} quantifies the association between *trial-level* treatment effects.

Intervals of ignorance and uncertainty fall into the frequentist paradigm. Alternatively, one could interpret the results from the sensitivity analysis in a Bayesian fashion as demonstrated in Sections 5.4 and 6.3. We could go further, and adopt a fully Bayesian framework. However, representing prior knowledge about the sensitivity parameters in a prior distribution can be difficult. For example, the correlation between the unidentifiable parameters in the prior distribution could impact the posterior distribution, but such correlations are extremely difficult to specify. Instead, representing prior knowledge in terms of simple ranges is easier. We therefore believe intervals of ignorance and uncertainty to be more useful concepts in this context.

Other models for the vector of potential outcomes can in principle be used. For example, Nevo and Gorfine (2022) propose a model where one correlation term models the cross-world dependence; hence, there is only one unidentifiable parameter, which simplifies the sensitivity analysis. However, such a model imposes *implicit* as-

sumptions on the unobservable association between cross-world potential outcomes. Roberts et al. (2024) start from a model with six correlated frailties and make several simplifying assumptions to aid in the identifiability. The remaining unidentifiable parameters—for which we proposed a formal sensitivity analysis—are fixed at preset values by Roberts et al. (2024).

One could also incorporate baseline covariates into the analysis to reduce uncertainty—both ignorance and imprecision—around the ICA. This idea has already been applied in the principal stratification framework (Long and Hudgens, 2013; Gilbert and Hudgens, 2008) and using baseline covariates in the ITCI framework is the subject of ongoing research. Specifically, the D-vine copula model can be formulated for the conditional distribution of $\tilde{\boldsymbol{Y}}$ given baseline covariates. If fitting such a model is computationally infeasible, simplified fitting strategies could be entertained such as a two-step approach, like the *inference for margins approach* (Joe, 2005). (This is also applicable to the model without baseline covariates.) Given that such a model can be fitted, the plausible set, Γ, would then relate to unidentifiable parameters of the conditional distribution of $\tilde{\boldsymbol{Y}}$ (e.g., the Spearman correlation between S_0 and S_1 *conditional on baseline covariates*); nonetheless, the sensitivity analysis can otherwise proceed as presented in Section 5.

7.2 Limitations

The findings emanating from a single-trial evaluation are strictly speaking only applicable within the population of the trial under evaluation. Extrapolating these results to future trials necessitates clinical and biological justifications beyond the available data. In contrast, the meta-analytic framework considers a population of clinical trials (Buyse et al., 2000b; Alonso et al., 2017). The extrapolation, or rather the generalization, comes naturally in that framework when the new trial can be considered as a random sample from that population of trials.

The proposed methods assume independent censoring; a violation of this unverifiable assumption may jeopardize the results. If independent censoring is believed to hold only conditionally on a set of available baseline covariates, the observable bivariate copula models could be fitted using inverse probability of censoring weights. The other steps remain the same. Alternatively, these bivariate models could be generalized along the lines of Czado and Van Keilegom (2023) to allow for dependent censoring.

Acknowledgements I. Van Keilegom gratefully acknowledges funding from the Research Foundation – Flanders and Fonds de la Recherche Scientifique – FNRS under the Excellence of Science (EOS) programme, project ASTeRISK [grant number 40007517]. F. Stijven gratefully acknowledges funding from Agentschap Innoveren & Ondernemen and Janssen through a Baekeland Mandate [grant number HBC.2022.0145]. The resources and services used in this work were provided by the Flemish Supercomputer Center, funded by the Research Foundation – Flanders and the Flemish Government.

Competing Interests

The authors have no competing interests to declare that are relevant to the content of this article.

Code and Data Availability

The *Surrogate* R package contains functions to (i) fit the D-vine copula model, (ii) perform the Monte Carlo sensitivity analysis, and (iii) compute the interval of ignorance and the uncertainty interval. The latest version of this package is available from `github.com/florianstijven/Surrogate-development`. This package is also available on CRAN.

The code used to analyze the data is available as a zipped R project, which also contains the advanced colorectal cancer data set. This project relies on the *renv* R package for managing the installed R packages, ensuring that the code can be rerun in a similar environment as ours. This R project can also be accessed via `github.com/florianstijven/-colorectal-surrogacy-analysis`.

References

Alonso A (2018) An information-theoretic approach for the evaluation of surrogate endpoints. In: Wiley StatsRef: Statistics Reference Online, John Wiley & Sons, Ltd

Alonso A, Van der Elst W, Molenberghs G, Buyse M, Burzykowski T (2015) On the relationship between the causal-inference and meta-analytic paradigms for the validation of surrogate endpoints. Biometrics 71(1):15–24

Alonso A, Van der Elst W, Molenberghs G, Buyse M, Burzykowski T (2016) An information-theoretic approach for the evaluation of surrogate endpoints based on causal inference. Biometrics 72(3):669–677

Alonso A, Bigirumurame T, Burzykowski T, Buyse M, Molenberghs G, Muchene L, Perualila N, Shkedy Z, Van der Elst W (2017) Applied surrogate endpoint evaluation methods with SAS and R. CRC Press, Boca Raton

Alonso Abad A, Ong F, Stijven F, Van der Elst W, Molenberghs G, Van Keilegom I, Verbeke G, Callegaro A (2024) An information-theoretic approach for the assessment of a continuous outcome as a surrogate for a binary true endpoint based on causal inference: application to vaccine evaluation. Statistics in Medicine 43(6):1083–1102

Burzykowski T, Molenberghs G, Buyse M, Geys H, Renard D (2001) Validation of surrogate end points in multiple randomized clinical trials with failure time end points. Journal of the Royal Statistical Society: Series C (Applied Statistics) 50(4):405–422

Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H (2000a) Statistical validation of surrogate endpoints: problems and proposals. Drug Information Journal 34(2):447–454

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

Conlon A, Taylor J, Elliott M (2017) Surrogacy assessment using principal stratification and a gaussian copula model. Statistical Methods in Medical Research 26(1):88–107

Corfu-A Study Group (1995) Phase III randomized study of two fluorouracil combinations with either interferon alfa-2a or leucovorin in advanced colorectal cancer. Journal of Clinical Oncology 13(4):921–928

Czado C (2019) Analyzing dependent data with vine copulas, vol 222. Springer, Cham

Czado C, Van Keilegom I (2023) Dependent censoring based on parametric copulas. Biometrika 110(3):721–738

Fine JP, Jiang H, Chappell R (2001) On semi-competing risks data. Biometrika 88(4):907–919

Fu H, Wang Y, Liu J, Kulkarni PM, Melemed AS (2013) Joint modeling of progression-free survival and overall survival by a bayesian normal induced copula estimation model. Statistics in Medicine 32(2):240–254

Georges P, Lamy AG, Nicolas E, Quibel G, Roncalli T (2001) Multivariate survival modelling: a unified approach with copulas. Available at SSRN: `https://ssrn.com/abstract=1032559`

Gilbert PB, Hudgens MG (2008) Evaluating candidate principal surrogate endpoints. Biometrics 64(4):1146–1154

Greco F, Figlin R, York M, Einhorn L, Schilsky R, Marshall E, Buys S, Froimtchuk M, Schuller J, Schuchter L, et al. (1996) Phase III randomized study to compare interferon alfa-2a in combination with fluorouracil versus fluorouracil alone in patients with advanced colorectal cancer. Journal of Clinical Oncology 14(10):2674–2681

Joe H (1989) Relative entropy measures of multivariate dependence. Journal of the American Statistical Association 84(405):157–164

Joe H (2005) Asymptotic efficiency of the two-stage estimation method for copula-based models. Journal of Multivariate Analysis 94(2):401–419

Kenward MG, Goetghebeur EJ, Molenberghs G (2001) Sensitivity analysis for incomplete categorical data. Statistical Modelling 1(1):31–48

Linfoot EH (1957) An informational measure of correlation. Information and Control 1(1):85–89

Long DM, Hudgens MG (2013) Sharpening bounds on principal effects with covariates. Biometrics 69(4):812–819

Meyvisch P, Alonso A, Van der Elst W, Molenberghs G (2020) On the relationship between association and surrogacy when both the surrogate and true endpoint are binary outcomes. Statistics in Medicine 39(26):3867–3878

Molenberghs G, Kenward M (2007) Missing data in clinical studies. John Wiley & Sons, Chichester

Molenberghs G, Kenward MG, Goetghebeur E (2001) Sensitivity analysis for incomplete contingency tables: the Slovenian plebiscite case. Journal of the Royal Statistical Society: Series C (Applied Statistics) 50(1):15–29

Nelsen RB (2006) An introduction to copulas, 2nd edn. Springer series in statistics, Springer, New York

Nevo D, Gorfine M (2022) Causal inference for semi-competing risks data. Biostatistics 23(4):1115–1132

Pankevich DE, Altevogt BM, Dunlop J, Gage FH, Hyman SE (2014) Improving and accelerating drug development for nervous system disorders. Neuron 84(3):546–553

Parast L, McDermott MM, Tian L (2016) Robust estimation of the proportion of treatment effect explained by surrogate marker information. Statistics in Medicine 35(10):1637–1653

Prentice RL (1989) Surrogate endpoints in clinical trials: definition and operational criteria. Statistics in Medicine 8(4):431–440

Roberts EK, Elliott MR, Taylor JM (2024) Surrogacy validation for time-to-event outcomes with illness-death frailty models. Biometrical Journal 66(1):1–20

Rosenbaum P, Rubin D (1983) The central role of the propensity score in observational studies for causal effects. Biometrika 70(1):41–55

Royston P, Parmar MK (2002) Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Statistics in Medicine 21(15):2175–2197

Vansteelandt S, Goetghebeur E, Kenward MG, Molenberghs G (2006) Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. Statistica Sinica 16(3):953–979

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