Assessment of the information theory approach to evaluating time-to-event surrogate and true endpoints in a meta-analytic setting

Pharmaceutical Statistics

Assessment of the Information Theory Approach to Evaluating Time-to-Event Surrogate and True Endpoints in a meta-analytic setting

Journal:	*Pharmaceutical Statistics*
Manuscript ID	PST-20-0049.R1
Wiley - Manuscript type:	Main Paper
Date Submitted by the Author:	28-Sep-2020
Complete List of Authors:	Dimier, Natalie; Roche Products Ltd, ; University of Reading, Mathematics and Statistics Todd, Susan; University of Reading, Mathematics and Statistics
Key Words:	Surrogate endpoint, Progression-Free Survival, Time-to-Progression, Information Theory
Abstract:	In many disease areas, commonly used long-term clinical endpoints are becoming increasingly difficult to implement due to long follow-up times and/or increased costs. Shorter-term surrogate endpoints are urgently needed to expedite drug development, the evaluation of which requires robust and reliable statistical methodology to drive meaningful clinical conclusions about the strength of relationship with the true long-term endpoint. This paper uses a simulation study to explore one such previously proposed method, based on information theory, for evaluation of time to event surrogate and long-term endpoints, including the first examination within a meta-analytic setting of multiple clinical trials with such endpoints. The performance of the information theory method is examined for various scenarios including different dependence structures, surrogate endpoints, censoring mechanisms, treatment effects, trial and sample sizes, and for surrogate and true endpoints with a natural time-ordering. Results allow us to conclude that, contrary to some findings in the literature, the approach provides estimates of surrogacy that may be substantially lower than the true relationship between surrogate and true endpoints, and rarely reach a level that would enable confidence in the strength of a given surrogate endpoint. As a result, care is needed in the assessment of time to event surrogate and true endpoints based only on this methodology.

Assessment of the Information Theory Approach to Evaluating Time-to-Event Surrogate and True Endpoints in a meta-analytic setting

Natalie Dimier[a,b]* and Susan Todd[b]

[a] Roche Products Ltd., Hexagon Place, 6 Falcon Way, Shire Park, Welwyn Garden City, AL7 1TW, UK.
[b] University of Reading, Department of Mathematics and Statistics, Whiteknights, PO Box 217, Reading, RG6 6AX, UK.
* Correspondence to: Natalie Dimier, Roche Products Ltd, Hexagon Place, 6 Falcon Way, Shire Park, Welwyn Garden City, AL7 1TW, UK.
Email: natalie.dimier@roche.com

Abstract

In many disease areas, commonly used long-term clinical endpoints are becoming increasingly difficult to implement due to long follow-up times and/or increased costs. Shorter-term surrogate endpoints are urgently needed to expedite drug development, the evaluation of which requires robust and reliable statistical methodology to drive meaningful clinical conclusions about the strength of relationship with the true long-term endpoint. This paper uses a simulation study to explore one such previously proposed method, based on information theory, for evaluation of time to event surrogate and long-term endpoints, including the first examination within a meta-analytic setting of multiple clinical trials with such endpoints.

The performance of the information theory method is examined for various scenarios including different dependence structures, surrogate endpoints, censoring mechanisms, treatment effects, trial and sample sizes, and for surrogate and true endpoints with a natural time-ordering. Results allow us to conclude that, contrary to some findings in the literature, the approach provides estimates of surrogacy that may be substantially lower than the true relationship between surrogate and true endpoints, and rarely reach a level that would enable confidence in the strength of a given surrogate endpoint. As a result, care is needed in the assessment of time to event surrogate and true endpoints based only on this methodology.

1. Introduction

A major component in the design of any clinical trial is the selection of a primary endpoint that reliably reflects clinical benefit for patients. Such endpoints are needed to estimate the necessary sample size and duration of a clinical trial, and are selected such that a positive result will lead to a beneficial impact to patients and healthcare systems. While many therapeutic areas have established gold standards for endpoint selection, these are being challenged by the increasing follow-up times required to reach endpoint maturity, increased costs and the need to provide effective treatments to patients in as short a time as possible. Such pressures have led to a sharp increase in research relating to endpoint choice, and notably in the area of surrogate endpoints. These short-term endpoints are designed to be used as replacements for gold standard clinical endpoints that are currently accepted by regulators, and as such require thorough investigation to ensure that they are meaningful and can predict long-term benefit for patients.

A number of statistical methodologies to assess the relationship between a potential surrogate endpoint and long-term (or 'true') endpoint have been proposed in the literature, including the first initial framework of Prentice (1989)[1], subsequent measures based on single trials[2] and meta-analytic measures.[3]-[10] A desirable feature of any methodology in this setting is that there is consistency in interpretation of the results across different proposed surrogate endpoint types, to allow comparison of different surrogate candidates. Furthermore, it is highly desirable that results and corresponding conclusions are consistent between different methodologies, to avoid a situation where different conclusions are drawn purely due to the choice of statistical methodology applied. Finally, it is imperative that results of any surrogacy evaluation accurately reflect the true underlying predictive ability of the surrogate endpoint under investigation. However, many of the methods currently available to assess potential surrogates have different assumptions and modelling structures depending on the endpoints under investigation[11], and comparability of surrogacy measures across these different techniques has not been established. The need for unified approaches that can incorporate many different endpoint types is therefore apparent, as is the need for comparative assessments across multiple proposed methodologies.

The most commonly used methodology for surrogacy assessment is the two-stage meta-analytic approach proposed initially by Buyse et al. (2000),[4], and extended by Burzykowski et al. (2001)[6] for time-to-event endpoints. This method has been used in a number of real-life applications, for example in leukaemia[12], follicular lymphoma[13] and other non-Hodgkins lymphomas[14]. However, as noted by Alonso and Molenberghs (2007)[10], the methodology has limitations in that it needs to be specifically adapted for different endpoint types and does not provide the unified interpretation that is highly desired. The approach also requires joint modelling of endpoints that can be complex. Further, Dimier and Todd (2017)[15] illustrated that the time to event version of the method, which is based on joint modelling of surrogate and true endpoints using copula models[6], can be adversely affected when used to assess endpoints that are not 'symmetrical', i.e. time to event endpoints that have a natural time ordering such that one cannot be longer than the other.

In a systematic review of surrogate endpoint methodology, Ensor et al. (2016)[16] recognise the need for unified approaches and noted that the information theory method proposed by Alonso and Molenberghs (2007)[10] has this advantage over the two-stage meta-analytic approach. The information theory method does not require definition of complicated joint distributions for surrogate and true endpoints in order to assess the relationship between them, and measures of surrogacy can be calculated using parameters that are routinely available in standard software packages and that are familiar to many statisticians. Importantly, the information theory approach does not make any assumptions of endpoint symmetry, and it has been speculated that it can be considered appropriate for the evaluation of time-ordered endpoints such as exploration of progression-free survival (time to disease progression or death) as a surrogate for overall survival (time to death) by considering the time difference between the surrogate and true endpoints as the true outcome of interest (Pryseley et al., 2011)[17]. This is considered of high importance in settings where long-term survival is becoming the primary challenge for timely conduct of new clinical trials.

The purpose of this paper is to investigate the information theory approach in the setting where results from several trials are available for assessing surrogacy (i.e. a meta-analytic context). This reflects the scenario most encountered in practice, and the most relevant for evaluating surrogacy over a variety of important clinical factors, such as mechanisms of action of different therapies and across different patient populations. The aim is to assess the performance of the method in estimating trial and individual level surrogacy in the setting of two time-to-event endpoints and for a variety of underlying strengths of surrogacy, sample sizes and surrogate endpoint candidates. While the endpoints are chosen to be reflective of oncology clinical trials, findings are applicable to all types of time-to-event endpoint. A number of new questions are being addressed by this research, namely how well the method performs when being applied in the important meta-analytic setting, including estimation of trial-level surrogacy in this setting, exploration of how it might be impacted by different dependency relationships between surrogate and true endpoints, and whether inclusion of information directly relating to the true long-term endpoint impacts the estimated strength of surrogacy. In Section 2 brief details of the information theory method designed for use with time-to-event endpoints are provided, followed in Section 3 by a description of a simulation study conducted to investigate the methodology. Finally results and discussion of identified challenges are given in Sections 4 and 5 respectively. Data sharing is not applicable to this article as no new data were created or analyzed in this study.

2. Information Theory Approach to Evaluating Surrogates

The value of a potential surrogate endpoint is commonly assessed using two measures; individual-level surrogacy and trial-level surrogacy [11]. Individual-level surrogacy measures the ability of a surrogate outcome to predict a long-term outcome for an individual patient, whereas trial-level surrogacy measures how well the unobserved treatment effect on the long-term endpoint can be predicted using the observed treatment effect on the surrogate endpoint. Both measures of surrogacy are considered important; while the individual-level provides information on the prognostic value of the surrogate outcome for an individual patient, the trial-level measure provides a critical evaluation of whether a strong treatment effect on the surrogate endpoint will lead to meaningful long-term clinical benefit at a population level. Both measures of surrogacy are of use to regulatory authorities responsible for making decisions based on the expectation of long-term benefit for patients.

Alonso and Molenberghs (2007)[10] first proposed the information theory measure as a way to evaluate both individual and trial level surrogacy, with key benefits including the applicability to many different endpoint types, and a unified interpretation of surrogacy measures across multiple surrogate endpoints. While the surrogacy measures based on this approach are calculated for single trials, a meta-analytic form can be estimated by taking a weighted average of the trial-level measures. The approach has been explored in the binary and continuous setting by Alonso and Molenberghs (2007)[10], and in the time-to-event setting by Pryseley et al. (2011) [17].

With the introduction of censored data that is expected in time to event endpoints, the work of O'Quigley & Flandre (2006)[18] suggests that the proposed information theory measure of individual-level surrogacy can be adversely impacted, and so Pryseley et al. (2011) [17] conducted further work to determine which of a selection of alternative measures could be used within the same framework to appropriately handle censored data. Results of this investigation suggest that one measure, proposed by Xu and O'Quigley (1999)[19] and based on a dependence measure for proportional hazards models, could be recommended for further use in evaluating individual-level surrogacy within an information theory context, maintaining the property of unified interpretability. The originally-proposed information theory based measure of trial-level surrogacy is unaffected by censoring, allowing it to be used immediately with time-to-event endpoints.

Suppose there exist data from a single clinical trial containing j=1,…,n patients, with surrogate and true outcomes denoted Sj and Tj respectively. To estimate individual-level surrogacy in this trial using the information theory approach, two Cox proportional hazards models are assumed, as follows:

1) $H_0 \colon \lambda_t(\mathrm{t} \,|\, \mathrm{Z}) = \lambda_{0t}(t)\exp\,(\beta_t \mathrm{Z}_j)$,
2) $H_1 \colon \lambda_{st}(\mathrm{t} \,|\, \mathrm{Z(t)}) = \lambda_{0st}(t)\exp\,(\,\beta_1 \mathrm{Z}_j + \beta_2 \mathrm{S}_j(t))$,

where $\lambda_{0t}(\mathrm{t})$ denotes the baseline hazard at time t, $\lambda_{0st}(\mathrm{t})$ denotes the baseline hazard at time t for a model adjusting for the surrogate outcome, β_t, β_1 and β_2 denote constant covariate coefficients, Z_j denotes the assigned treatment group and $Z(t)$ denotes a vector of time-dependent covariates that includes treatment group. Only covariates of treatment and surrogate outcome will be discussed herein; the addition of other covariates does not impact on the procedure for calculation of the surrogacy measures. In model H_1, the covariate representing the surrogate endpoint is reflected as a time-varying value, to denote whether the surrogate event has occurred or not at time t.

In this model representation, H_0 corresponds to a model where β_2 is assumed to be zero, such that the surrogate has no effect on T and surrogacy is poor. The 'alternative' model, H_1, corresponds to a model where β_2 has no restriction, thereby providing the 'information gain' that comes from inclusion of the surrogate outcome, S, in the model. Quantification of the information difference between these two models therefore indicates the strength of prediction of outcome T that can be gained through knowledge of the surrogate outcome, S. Using the notation of Pryseley et al. (2011) [17], the recommended measure of individual-level surrogacy for the trial is defined as

$$R^2_{XOQ} = 1 - \exp\,(-\Gamma(\mathrm{H1,H0};\beta)),$$

where β represents the true values of (β_1, β_2) and

$$\Gamma(\mathrm{H1,H0};\beta) = 2\int_T\int_Z \log\left[\frac{g(z|t;\beta)}{g(z|t;\hat{\beta})}\right] g(z\,\big|\,t;\beta)\;dzdF(t),$$

where the parameter $\hat{\beta}$ denotes the value maximising the likelihood of model H_0 with respect to β, g(z|t,β) denotes the conditional distribution of Z|T and F(t) denotes the marginal distribution function of T. In practice, the Kaplan-Meier function is proposed (Kaplan and Meier, 1958)[20] as the marginal distribution F(t), and for the conditional distribution of Z|T at each event time, Xu and O'Quigley (1999) [19] propose to use the conditional probability of patient j having an event at time t given their covariate values at that time, based on the partial likelihood of the Cox proportional hazards models (Cox, 1972)[21]. Xu and O'Quigley (1999) [19] note that the R^2_{XOQ} measure is dependent on the total duration of follow-up of the trial, and propose an adjusted version that accounts for this, dividing by the total period of follow-up.

This value of R^2_{XOQ} can be calculated using quantities estimated from standard software packages, making it very appealing, and confidence intervals can be constructed by re-calculating the R^2_{XOQ} measure using the confidence limits for the surrogate covariate coefficient using the same approach. In the meta-analytic setting, R^2_{XOQ} can be calculated for each trial and a weighted estimate used to provide an overall measure.

For the time to event setting, Pryseley et al. (2011) [17] noted that the information theory approach can also be immediately applied to time-ordered endpoints, where S<=T, for example when using overall survival as the true endpoint where all potential surrogates are restricted to occur prior to T. Instead of modelling the impact of the surrogate outcome directly on the true outcome in the Cox proportional hazards models, the time between these two endpoints can be used as the outcome measure. For example, in the context being described herein, use of post-progression survival as a time-to-event endpoint can be used in place of T, with the status of the surrogate outcome (event or not) being used as the covariate in model H_1.

At the trial-level, multiple trials are needed to evaluate the relationship between treatment effects on S and T, and so a meta-analysis is required. In the information theory framework, the same concepts can be applied to trial-level surrogacy in estimating how much of the uncertainty in treatment effect on the true endpoint can be reduced through knowledge of the treatment effect on the surrogate endpoint. Alonso and Molenberghs (2007)[10] note that under the assumption of a linear relationship between treatment effects on S and T, the square of the correlation coefficient between treatment effects immediately has an interpretation from the information theory perspective. Therefore, the trial-level surrogacy, denoted R^2_{trial}, can be estimated by taking the square of the correlation coefficient between treatment effects on T from model H_0 and the treatment effect on S estimated from the following model:

1) $H_s\colon \lambda_s(\mathrm{s}\,|\,\mathrm{Z}) = \lambda_{0s}(\mathrm{s})\exp\left(\beta_S \mathrm{Z}_j\right)$.

As before, $\lambda_{0s}(\mathrm{s})$ denotes the baseline hazard at time s, and β_s denotes a constant covariate coefficient for treatment, Z_j.

The only published evaluation of the information theory approach in the context of time-to-event endpoints is that of Pryseley et al. (2011) [17], who conducted a simulation study to assess individual-level surrogacy for a time-to-progression endpoint (TTP: defined as time from study entry to the earliest sign of disease progression) as a surrogate for overall survival (OS: defined as time from study entry to death). This study considered individual trials containing 100-2000 patients, with no censoring and with 20-70% censoring, and for various true underlying individual-level surrogacy levels.

Since the aforementioned investigation was based on data generated from only a single clinical trial, no consideration of the meta-analytic setting, nor of trial-level surrogacy, has been conducted in the literature for time to event endpoints. While research has been published for other endpoint types, estimation of treatment effects on time to event endpoints can be dependent on the level of censoring present in the data, and so it is not immediate that results of investigations from other endpoint types can be extrapolated to the time-to-event setting. The question of how well the information theory approach can perform in a meta-analysis of time to event data therefore remains unanswered and is the primary goal of the research presented herein. Secondly, it is also of interest to assess whether the information theory method is sensitive to the underlying dependence structure of the surrogate and true endpoints within the clinical trials, which has been shown to cause issues for other surrogacy evaluation approaches[15]. Finally, it is of interest to explore the suggestion that the information theory approach can be used to reliably assess surrogacy of time-ordered endpoints, such as surrogates for overall survival. If so, this would be considered of extremely high value in settings where long survival times are hampering efforts to conduct new clinical trials of experimental therapies. Each of these research questions will be explored through the use of a simulation study, details of which are presented in Section 3.

3. Simulation Study

A key consideration in the set-up of a simulation study to assess any surrogacy evaluation method is how the underlying surrogacy (trial and individual) can be adequately controlled. Ideally, the parameter being estimated by the surrogacy approach would be directly controlled. However, the parameters that are required in calculation of the information theory method of individual-level surrogacy make this very difficult, as they are estimated from conditional models using the likelihood ratio, meaning that each sample would have a slightly different underlying value.

Previous investigations of the R^2_{XOQ} measure have been based on datasets simulated according to a Cox proportional hazards model (Xu and O'Quigley, 1999) [19], where a single covariate was included in the model and the respective covariate coefficient was used to control the overall strength of association with the outcome variable. Whilst this allows for specification of the strength of this relationship, there could be considerable subjectivity in the selection of coefficient values, and the impact of this on the resulting estimation of R^2_{XOQ} is not currently clear. Additionally, the strength of covariate coefficient, or hazard ratio, may be interpreted very differently depending on the disease context.

Instead, the study of Pryseley et al. (2011) [17] was based on data generated using a Clayton copula function, controlling association between endpoints through the copula parameter, τ, which is designed to reflect the strength of association between S and T based on the chosen dependence structure. Whilst the value of τ may not perfectly reflect the true information theoretic measure of association, such an approach allows for overall control of individual (τ) and trial (R^2_{trial}) association levels, subject to sample variability, with a clear interpretation (strength of association between S and T after adjusting for treatment and trial effects) and scale (ranging 0-1). In order to be consistent with the approach taken by Pryseley et al. (2011) [17], and to allow meaningful comparison across these simulation studies, data generation herein was also conducted using this copula model. To assess the robustness of the results to this approach, two different copula models were used to change the underlying dependence structure and evaluate whether this has any impact on the results. Further selected scenarios were also generated using an algorithm based on the lognormal distribution, that does not use a copula model; these scenarios are displayed in Table 1. A brief description of the simulation setup is provided below.

In this study, a true endpoint of OS (defined as time from study entry to death) is of interest, to reflect the most challenging setting where surrogate endpoints are most needed. Surrogate endpoints of time to progression (TTP: defined as time from study entry to disease progression) and progression-free survival (PFS: defined as time from study entry to the earliest of disease progression or death) are considered, with the goal to evaluate how inclusion of data from the true endpoint into the surrogate definition might impact the performance of the methodology.

In all data generation methods, model parameters are chosen to reflect a scenario where the median value of the surrogate (5-6 months) is approximately half of that of the true endpoint (11-12 months). Treatment effects are chosen such that the effect on S (hazard ratio ~0.67) is slightly stronger than that on T (hazard ratio ~0.82), to reflect the potential influence of post-progression therapies on the observed treatment effect on OS. Censoring is applied by drawing an exponential random variable and comparing to the simulated event values, scaling the random value to control the proportion of censoring in the data. Since the true endpoint of interest is OS, the value of TTP as the surrogate is also censored by the true endpoint, if it occurs first. For PFS, when death occurs prior to progression, the patient is considered to have an event at the time of death, and additional censoring is not applied. In order to investigate the potential for false positive or false negative results, the strength of individual and trial-level surrogacy was varied from very low (0.2, to reflect a true lack of surrogacy) to very high (0.8, to reflect a very strong level of surrogacy).

Data Generation

Data generation using the Clayton copula was conducted using the conditional distribution method (Burzykowski et al. 2001)[6]. This algorithm draws two independent random variables from a Uniform(0, 1) distribution, which are then transformed to be distributed according to the joint survival function defined by the Clayton copula function, with strength of dependence controlled using the copula dependence parameter, τ. Once transformed, the two uniform random variables have the required shape and strength of association and can be further transformed to survival outcomes according to the desired marginal survivor functions. Based on these marginal functions, the joint survival function provides strong upper-tail dependence and weaker lower-tail dependence (see Burzykowski (2001)[6] for details). For consistency with previous studies, exponential marginal distributions were selected.

To generate data according to the dependence structure of the Gumbel copula, which provides stronger lower-tail dependence (i.e. at earlier event times) and weaker upper-tail dependence, the mixtures of powers algorithm described by Trivedi and Zimmer (2007)[22] was used, with parameters selected to ensure comparability of overall summary characteristics with the data generated using the Clayton copula.

In order to evaluate whether the use of copula functions may bias the data generation, lognormally-distributed data were also generated using an algorithm previously used in the context of comparing estimators of Kendall's τ (Hsieh, 2010)[23]. Using this method, the parameter that controls the association between endpoints remains as Kendall's τ, but the value is transformed for use within a covariance matrix of a bivariate normal distribution to control the strength of correlation between endpoints. Although this method also makes use of the joint distribution between the two endpoints, it is not reliant on the choice of copula.

To generate individual values of surrogate and true outcomes using the lognormal data generation method, a three-step process was created. First, two variables, one for each of the surrogate and true outcome, were generated from a bivariate Normal distribution with a respective correlation coefficient ρ chosen to reflect Kendall's τ through the relationship $\rho = \sin\frac{\tau\pi}{2}$; (Kruskal, 1958)[24]. These values were then rescaled to ensure that the final distributions of the generated times are similar to those generated using the copula models; since Kendall's τ is based on ranks of variables rather than specific values, the underlying value of T is unaffected by such a monotonic transformation. Finally, the values were exponentiated to obtain lognormally distributed time-to-event values, and the effect of treatment and the underlying trial-level association were incorporated by scaling the generated time-to-event value by the required random effects.

Simulations were run on a Windows 7 64-bit machine with 4 GB RAM, using macros based on SAS software, version 9 for Windows[25].

4. Results

4.1 Individual-Level Surrogacy

In this study, individual-level surrogacy was measured within each trial and an overall estimate for the meta-analytic setting was taken as a weighted average across trials, using the trial sample size as the weight. The notation R^2_{XOQ} here denotes the weighted meta-analytic estimate. Since the simulations were designed based on a fixed sample size and proportion of censoring, it is expected that weighting by the number of observed OS events would lead to comparable conclusions.

Figure 1 illustrates the estimated R^2_{XOQ} values across the scenarios of interest when there are six trials available for analysis, each containing 120 patients. Each boxplot shows the range of estimated values across all simulation runs, with the level of censoring and the underlying individual-level association along the x-axis. Within the figure, the individual plots display results from the TTP endpoint on the top row and PFS endpoint on the bottom row. Horizontal dashed lines at y= 0.2, 0.5 and 0.8 represent the true individual-level surrogacy being estimated by each set of boxplots.

Across all data generation methods, results based on TTP as the surrogate endpoint (top row) demonstrate that the information theory method produces estimates of R^2_{XOQ} that are lower than the value of τ used in data generation, and this is true across all levels of individual association, and for all proportions of censoring. Encouragingly, the estimates increase as τ increases, suggesting that whilst they are far from the input value, they do reflect increasing magnitudes of association. In further exploration of underlying association levels as high as 0.9, the estimated results had a median of approximately 0.45 and never exceeded 0.6, reflecting levels that would remain too low to provide sufficient confidence that a surrogate endpoint could be used in future clinical trials. However, the main concern is the very large ranges of results, which increase as the true level of association increases. Estimates of medium (τ = 0.5) and high (τ = 0.8) strengths of association are widely spread, which hampers interpretation. Coupled with the overall lower estimates of individual-level association, it is unlikely that even the strongest surrogates could be identified, with estimates of R^2_{XOQ} ranging from approximately 0.2 to 0.7 when the true association is strong. Importantly, the estimated measures appear to be generally robust to the proportion of censoring in the data, which provides confidence that datasets with varying amounts of censoring and durations of follow-up can provide consistent results. Further, the measure appears to be robust to the method of data generation, suggesting that the method is not impacted through the shape of the dependence structure between S and T.

Results based on PFS as the surrogate endpoint (bottom row) highlight strong similarities to those for TTP, indicating that there is little impact from endpoints that are not symmetrical or have a natural time-ordering. Across the range of simulation scenarios examined, estimates of R^2_{XOQ} are slightly higher than those for TTP, which could be expected based on the definition of progression-free survival incorporating information related to the true endpoint. Unfortunately, the high variability remains, meaning that even for the highest underlying association level of 0.8 the estimated value could be as low as approximately 0.4, which could convince clinicians and regulatory bodies that the surrogate was not worthy of further consideration.

Figure 1 Estimated values of R^2_{XOQ} (N=6, n=120)

To determine whether more data can improve estimation and reduce variability, further simulations were conducted using a total of ten trials, each containing 500 patients. Given the similarity of results between the Gumbel copula and the lognormal approach in the smaller sample sizes, only Clayton and lognormal generated data were used to evaluate these larger samples, which can be found in Figure 2.

Figure 2 Estimated values of R^2_{XOQ} (N=10, n=500)

For both TTP and PFS scenarios, the availability of larger clinical trial databases has improved estimation, specifically with respect to the variability in estimated values of R^2_{XOQ}. The wide ranges of estimates observed with smaller sample sizes was considered to hinder the interpretation of the results, whereas these additional scenarios demonstrate reasonably similar estimates across all simulation runs. While variability increases slightly with censoring, the results remain interpretable and there is a reduction in the overlap of estimates between true underlying surrogacy values, providing confidence that the proportion of censoring in the data would not change the conclusions of the analysis. However, it remains evident that even under truly strong association, the method cannot reach values of R^2_{XOQ} greater than approximately 0.6 for TTP and 0.8 for PFS. Therefore, while improved variability can be achieved through increased data, the underestimation remains of concern and promising surrogates may still be overlooked.

4.2 Trial-Level Surrogacy

Figure 3 contains similar boxplots to those for individual-level surrogacy in Figure 1, with the y-axis now representing trial-level surrogacy. Overall, across all simulation scenarios, estimates of R^2_{trial} were highly variable. Based on TTP (top row), there is a slight upwards trend as the true R^2_{trial} increases, however this increase is negligible in comparison to the large variability, and does not allow for reliable conclusions. Where there is severe over- or under-estimation observed for the TTP scenarios, estimates of R^2_{trial} based on PFS as the surrogate endpoint are generally overestimated. Across almost all settings, estimates range across the entire [0, 1] interval, with median values generally lying in the range of 0.2 to 0.5 for TTP and 0.5 to 0.8 for PFS, and there is little that can be concluded from these results. Overall, the information theory method cannot be deemed appropriate for use in the setting of small trial numbers, as has been observed for other surrogacy approaches[15].

Figure 3 Estimated values of R^2_{trial} (N=6, n=120)

As for individual-level surrogacy, a selection of additional simulations were conducted using a total of ten trials, each containing 500 patients, to determine whether more data can improve estimation. These results can be found in Figure 4, and illustrate that increasing the number of trials from six to ten and increasing trial sizes from 120 to 500 patients offers some improvement in that the variability of the estimates is reduced. However, across the majority of the scenarios, the estimates remain far from the underlying value used in simulation and have highly overlapping distributions. While the slight upwards trend as the true R^2_{trial} increases remains, the high variability and bias in estimation hampers interpretation in a real-life setting.

Figure 4 Estimated values of R^2_{trial} (N=10, n=500)

4.3 Time-ordered endpoints

The paper by Pryseley et al. (2011) [17] speculates that the information theory method for time-to-event endpoints can naturally be applied to time-to-event endpoints that have a natural ordering, such as when overall survival is the true outcome of interest. Since this has not yet been examined in the literature, a selection of scenarios generated for the simulation study described here were used to test this assumption by taking the outcome variable to be post-progression survival (T −S), with the surrogate outcome accounted for through describing the progression status (progression [1] or not [0]) as a binary covariate.

Figure 5 illustrates the estimated R^2_{XOQ} values for the time-ordered approach, again containing six trials, each containing 120 patients, with 0-30% censoring shown on the x-axis, TTP on the top row, PFS on the bottom row, and each level of underlying surrogacy highlighted.

Across all the scenarios examined, it was found that the proposed adjustment using T-S as the outcome variable leads to extremely poor estimation of R^2_{XOQ}. Regardless of whether the surrogate endpoint was TTP or PFS or whether the Clayton and Gumbel-generated data were considered, there was no impact on the results, with values of R^2_{XOQ} rarely exceeding a value of 0.2 even for the largest values of τ.

Figure 5 Estimated values of R^2_{XOQ} for time-ordered endpoints (N=6, n=120)

The most likely cause of the poor performance in this context is that key information relating to the surrogate outcome is ignored. In reality, the surrogate outcome includes not only the disease status, but also the time at which the disease status changed. The outcome (T-S) does not capture the time of disease progression, only the duration of time after progression that a patient survived, and so it could be expected that it is not sensitive enough to reliably quantify the relationship between disease status and true outcome. It is worth noting that the information theory approach is not based on joint modelling of endpoints, and so does not assume any endpoint symmetry. The attempt to correct for a problem that does not exist is therefore unnecessary, and should not be recommended. Future use of the information theory method should therefore maintain the outcome variable T and allow for the assessment of time-ordering of endpoints through the use of a time-dependent covariate to represent S.

For completeness, estimation of R^2_{trial} based on the time-ordered endpoint structure can be found in the Supplementary Results.

5. Discussion

The main aim of this research was to assess the performance of the information theory approach in evaluating surrogacy within a meta-analysis based on time-to-event endpoints. We believe this to be the only detailed assessment of the approach in this setting, allowing estimation of both individual and trial level surrogacy when based on data from multiple clinical trials. It is also the first assessment of the approach with surrogate endpoints that also incorporate data from the true clinical endpoint. Finally, it is the only known investigation of the approach in a setting of time-ordered endpoints, where post-progression survival is used as the outcome measure to account for the fact that the surrogate endpoint must always occur prior to the true endpoint.

Based on the numbers of trials included in this study (six or ten), the information theory approach was unable to reliably detect the underlying strength of trial level association, and in the worst case could lead to false positive results that would incorrectly support an endpoint being considered a sufficient surrogate for use in future clinical trials. Minor improvements were observed with the larger number of trials, however these were insufficient to recommend the method for use. In general, the results described herein demonstrate that the evaluation of surrogate endpoints is very challenging when the number of clinical trials is limited, and supports the need for collaborative efforts between the pharmaceutical industry and academic institutions to ensure that meaningful data can be collated for such research purposes.

At the individual-level, the results have demonstrated that the information theory approach consistently under-estimates the underlying surrogacy strength, regardless of the endpoint, data generation method, level of censoring in the dataset or sample size. That said, the approach showed very encouraging results with respect to insensitivity to the proportion of censoring in the datasets, the dependence structure between S and T and the type of surrogate endpoint. However, despite the potential benefits of being easy to estimate and having a unified interpretation, the results herein have demonstrated that the information theory method is insufficiently accurate to draw meaningful conclusions when applied to small-sample meta-analyses with time to event surrogate and true endpoints. Based on recent examples[13,14], trial-level surrogacy values of at least 0.8 are considered necessary to truly establish a surrogate endpoint for future use as a primary endpoint for regulatory and healthcare agency decision making. If such a high threshold were also set for individual-level surrogacy then it is highly unlikely that the information theory approach would lead to regulatory endorsement of any surrogate endpoint, regardless of the underlying strength of association. It should be highlighted that the surrogate endpoints selected for investigation in this research are examples reflecting commonly used endpoints in oncology clinical trial research. While results are expected to translate to time-to-event surrogate endpoints in general, the biological rationale based on the disease under study, and the potential for confounding of results through selection of specific patient populations or treatment mechanisms, should also be strongly considered when considering the plausibility of any surrogate endpoint.

It should be noted that results and conclusions of this research differ to some extent from those of Pryseley et al. (2011) [17], who concluded that the method performed acceptably when the percentage of censored observations on the surrogate by the true endpoint is not too high. Upon further investigation of the software used for their study (made available by the author), a small discrepancy in the data generation formula was identified, meaning that the data were not simulated according to the intended strength of association and the reference value of τ used for bias calculations was subsequently incorrect. In testing of programming with this anomaly, results reported by Pryseley et al. (2011)[17] were able to be replicated, suggesting that this is the primary cause of the differences in observed results. Importantly, both simulation studies are consistent in that findings demonstrate downward bias for very high values of individual-level surrogacy.

The main limitation of the simulation study described here has been noted previously; the true underlying individual-level surrogacy cannot be controlled via simulation. Since the R^2_{XOQ} parameter is calculated from conditional models and is based on a likelihood ratio, each generated sample can have a slightly different value. An alternative representation of the individual-level surrogacy is therefore needed to ensure that each sample is being compared to an intended level of association between endpoints, and for this purpose Kendall's τ was used. The impact of this is that bias estimates may be incorrectly over- or under-estimated. However, three different data generation algorithms were explored, and the results were broadly consistent across all scenarios investigated, therefore the findings are considered to be robust and it is reasonable to conclude that the information theory method as currently proposed risks missing truly promising surrogate endpoints.

A further limitation of the simulation study is that only one set of treatment effects were examined (HR=0.67 for PFS and 0.82 for OS for all trials). In the data generation procedure, a change in treatment effect has no impact on the underlying strength of association, however it could be possible that a change in treatment effect could cause a difference to the Cox model parameters that are estimated within the information theory approach. Stronger treatment effects were therefore also considered as a sensitivity analysis for selected scenarios using Clayton-copula generated data (HR=0.50 for PFS and 0.67 for OS) and demonstrate findings that are highly consistent with the originally selected treatment effects, with the median and ranges of estimates of R^2_{XOQ} being very similar across all levels of τ (see Supplementary Material). Hence, the selection of specific treatment effects is not considered to have confounded the results of the simulation study.

Finally, the information theory approach is based on an assumption of proportional hazards, such that Cox models can be used to estimate treatment and surrogate covariate coefficients. The data generation procedure forced proportional hazards through implementation of a time constant treatment effect, and there was no consideration of the impact on modelling when this assumption was violated. Since the Cox model can be used with time-dependent covariates, it is possible to adjust for some forms of non-proportional hazards, but such settings were not investigated in this study. Examination of non-proportional hazards was conducted by Pryseley (2009)[26], who concluded that the measure R^2_{XOQ} in a single-trial setting performed acceptably well when the proportion of censoring was low to moderate, however further work could be considered to understand the extent of violation that must be observed for the information theory approach to show evidence of significantly deteriorated performance.

It should be noted that the method investigated in this study, along with many others, makes the assumption that the surrogate endpoint is in a single, direct causal pathway of the disease under study and the ultimate true clinical outcome, and that all of the treatment effect on the true outcome is mediated through the surrogate endpoint. Such assumptions are unverifiable, and development of methodology to estimate the causal effect of surrogates and reduce confounding of either alternative disease pathways, or alternative impacts of treatment effect on surrogate and true endpoints, will be of increasing importance for future research (Frangakis and Rubin, 2002).

Overall, results of this simulation study allow us to conclude that the information theory approach struggles to reach a level that would enable confidence in the strength of a given surrogate endpoint. As a result, caution should be used when applying the information theory approach to evaluating surrogate endpoints of a time-to-event nature, primarily because the estimates of surrogacy may be substantially lower than the true relationship between surrogate and true endpoints.

References

1. Prentice R. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine 1989*; **8**:431–440.
2. Buyse M, Molenberghs G. Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics 1998;* **54**:1014–1029.
3. Daniels MJ, Hughes MD. Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine 1997*; **16**:1965–1982.
4. Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H. The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics 2000;* **1**:49–67.
5. Gail MH, Pfeiffer R, van Houwelingen HC, Carroll RJ. On meta-analytic assessment of surrogate outcomes. *Biostatistics 2000;* **1**:231–246.
6. Burzykowski T, Molenberghs G, Buyse M, Geys H, Renard D. Validation of surrogate endpoints in multiple randomized clinical trials with failure time endpoints. *Applied Statistics 2001;* **50**:405–422.
7. Tibaldi F, Barbosa FT, Molenberghs G. Modelling associations between time-to-event responses in pilot cancer clinical trials using a Plackett-D model. *Statistics in Medicine 2004*; **23**:2173–2186.
8. Alonso A, Molenberghs G, Burzykowski T, Renard D, Geys H, Shkedy Z, Tibaldi F, Cortinas Abrahantes J, Buyse M. Prentice's Approach and the Meta-Analytic Paradigm: A Reflection on the Role of Statistics in the Evaluation of Surrogate Endpoints. *Biometrics 2004*; **60**:724–728
9. Burzykowski T, Buyse M. Surrogate threshold effect: An alternative measure for meta-analytic surrogate endpoint validation. *Pharmaceutical Statistics 2006;* **5**:173–186.
10. Alonso A, Molenberghs G. Surrogate Marker Evaluation from an Information Theory Perspective. *Biometrics 2007*; **63**:180–186.
11. Burzykowski T, Molenberghs G, Buyse M. The Evaluation of Surrogate Endpoints. Springer: New York, 2005.
12. Buyse M, Michiels S, Squifflet P, Lucchesi KJ, Hellstrand K, Brune ML, Castaigne S, Rowe JM. Leukemia-free survival as a surrogate end point for overall survival in the evaluation of maintenance therapy for patients with acute myeloid leukemia in complete remission. *Haematologica 2011*; **96**:1106-12.
13. Shi Q, Flowers CA, Hiddemann W, Marcus R, Herold M, Hagenbeek A, Kimby E, Hochster H, Vitolo U, Peterson BA, Gyan E, Ghielmini M, Nielsen T, De Bedout S, Fu T, Valente N, Fowler NH, Hoster E, Ladetto M, Morschhauser F, Zucca E, Salles G, Sargent DJ. Thirty-Month Complete Response as a Surrogate End Point in First-Line Follicular Lymphoma Therapy: An Individual Patient-Level Analysis of Multiple Randomized Trials. *Journal of Clinical Oncology 2017;* **35**: 552-560

14. Shi Q, Schmitz N, Ou F-S, Dixon J, Cunningham D, Pfreundschuh M, Seymour JF, Jaeger U, Habermann TM, Haioun C, Tilly H, Ghesquieres H, Merli F, Ziepert M, Herbrecht R, Flament J, Fu T, Coiffier B, Flowers CR. Progression-Free Survival as a Surrogate End Point for Overall Survival in First-Line Diffuse Large B-Cell Lymphoma: An Individual Patient–Level Analysis of Multiple Randomized Trials (SEAL). *Journal of Clinical Oncology 2018*; **36**: 2593-2602
15. Dimier N, Todd, S. An investigation into the two-stage meta-analytic copula modelling approach for evaluating time-to-event surrogate endpoints which comprise of one or more events of interest. Pharmaceutical Statistics 2017; **16**: 322-333.
16. Ensor H, Lee RJ, Sudlow C, Weir CJ. Statistical approaches for evaluating surrogate outcomes in clinical trials: A systematic review. *Journal of Biopharmaceutical Statistics 2016*; **26:** 859-879
17. Pryseley A, Tilahun A, Alonso A, Molenberghs G. An information-theoretic approach to surrogate-marker evaluation with failure time endpoints. *Lifetime Data Analysis 2011*; **17**:195–214.
18. O'Quigley J, Flandre P. Quantification of the Prentice criteria for surrogate endpoints. *Biometrics 2006*, **62**:297–300.
19. Xu R, O'Quigley J. A R^2 type measure of dependence for proportional hazards models. Journal of Nonparametric Statistics 1999; **12:**83–107.
20. Kaplan EL, Meier P. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association 1958*, **53**:457–481.
21. Cox DR. Regression models and life-tables. *Journal of the Royal Statistical Society 1972*, **34**:187–202.
22. Trivedi PK, Zimmer DM. Copula modeling: An introduction for practitioners. *Foundations and Trends in Econometrics 2007*, **1**:1–111.
23. Hsieh J-J. Estimation of Kendall's tau from censored data. *Computational Statistics and Data Analysis 2010;* **54**:1613–1621
24. Kruskal WH. Ordinal Measures of Association. *Journal of the American Statistical Association 1958*; **53**:814–861.
25. Copyright, SAS Institute Inc. Cary, NC, USA
26. Pryseley A. Marker methodology, with focus on time-to-event outcomes. *Unpublished Ph.D. dissertation, Hasselt University, 2009.*
27. Frangakis CE, Rubin DB. Principal Stratification in Causal Inference. *Biometrics 2002* **58**: 21-29.

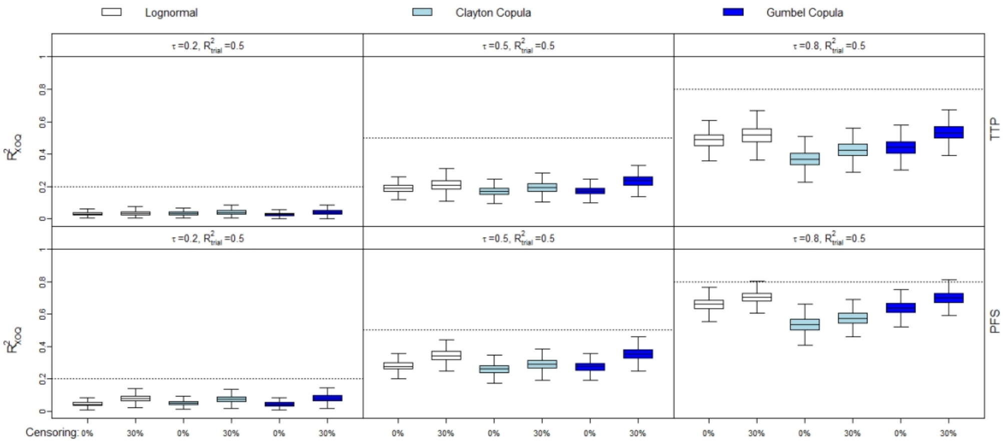
Table 1: Simulation Scenarios

Factor	**Scenarios under simulation**
Surrogate Endpoint	Time-to-Progression (TTP), Progression-Free-Survival (PFS)
True Endpoint	Overall Survival (OS), post-progression survival (for the time-ordered endpoint only)
Data Generation[1]	Clayton, Gumbel, Lognormal
Number of trials	6, 10
Number of patients per trial	120, 500
Trial-level association[2]	0.2, 0.5, 0.8
Individual-level association[3]	0.2, 0.5, 0.8
Censoring Rate (on T)	0%, 30%

[1] Initial simulations considered three data generation methods using 6 trials each containing 120 patients. Further examination was conducted using 10 trials of 500 patients for the Clayton and Lognormal methods only.

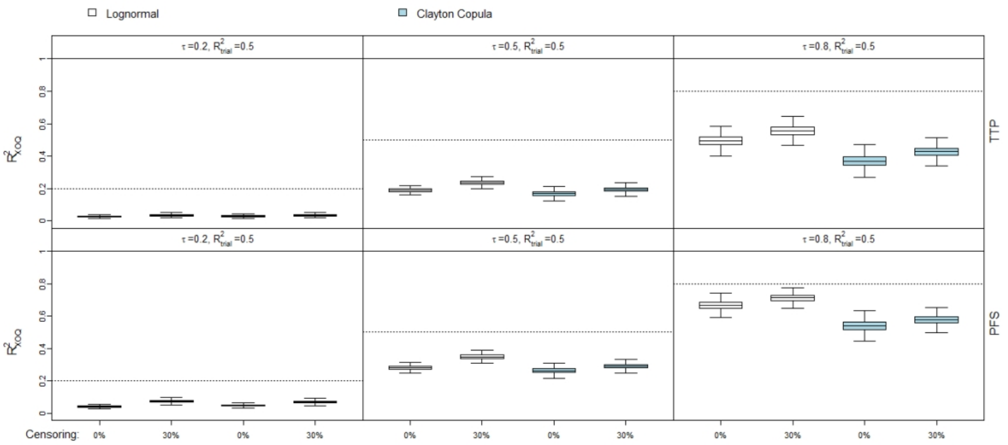
[2] When evaluating trial-level surrogacy, individual-level surrogacy is fixed at a value of 0.5.

[3] When evaluating individual-level surrogacy, trial-level surrogacy is fixed at a value of 0.5.



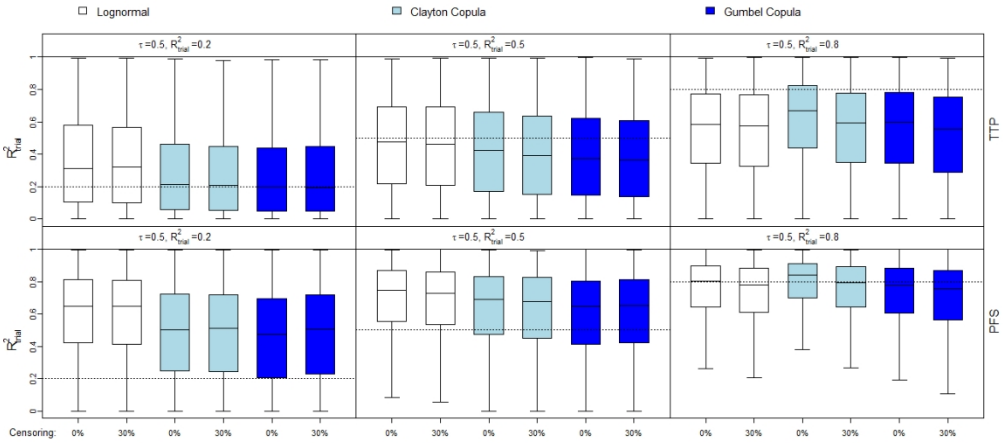
Estimated values of R^2_{XOQ} (N=6, n=120)

463x211mm (72 x 72 DPI)



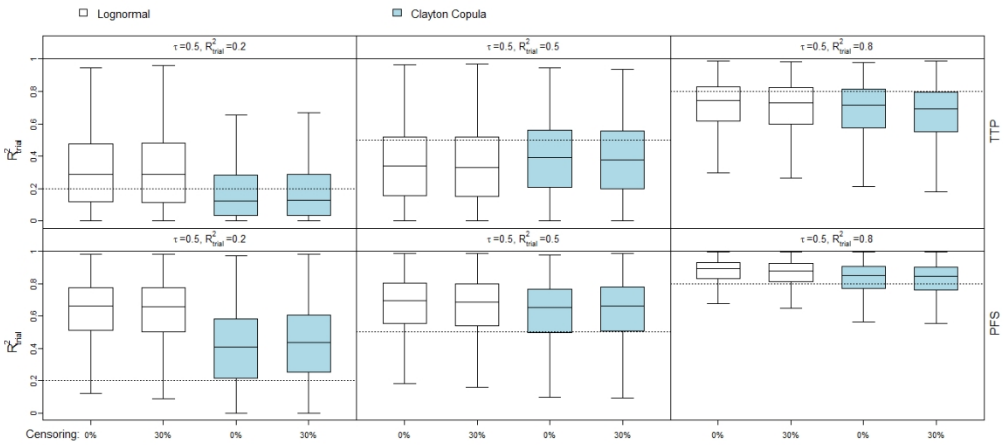
Estimated values of R^2_{XOQ} (N=10, n=500)

463x211mm (72 x 72 DPI)



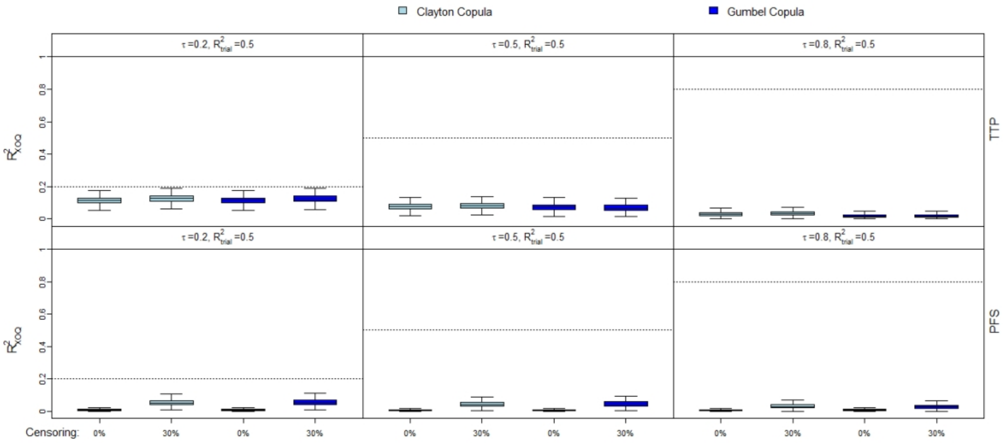
Estimated values of R^2_{trial} (N=6, n=120)

463x211mm (72 x 72 DPI)



Estimated values of R^2_{trial} (N=10, n=500)

463x211mm (72 x 72 DPI)



Estimated values of R^2_{XOQ} for time-ordered endpoints (N=6, n=120)

463x211mm (72 x 72 DPI)