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

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

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

Clinical trials of experimental treatments must be designed with primary endpoints that directly measure clinical benefit for patients. In many disease areas, the recognised gold standard primary endpoint can take many years to mature, leading to challenges in the conduct and quality of clinical studies. There is increasing interest in using shorter-term surrogate endpoints as substitutes for costly long-term clinical trial endpoints; such surrogates need to be selected according to biological plausibility, as well as the ability to reliably predict the unobserved treatment effect on the long-term endpoint. A number of statistical methods to evaluate this prediction have been proposed; this paper uses a simulation study to explore one such method in the context of time-to-event surrogates for a time-to-event true endpoint. This two-stage meta-analytic copula method has been extensively studied for time-to-event surrogate endpoints with one event of interest, but thus far has not been explored for the assessment of surrogates which have multiple events of interest, such as those incorporating information directly from the true clinical endpoint. We assess the sensitivity of the method to various factors including strength of association between endpoints, the quantity of data available and the effect of censoring. In particular, we consider scenarios where there exist very little data on which to assess surrogacy. Results show that the two-stage meta-analytic copula method performs well under certain circumstances and could be considered useful in practice, but demonstrates limitations that may prevent universal use.

Keywords: Surrogate Endpoint; Meta-Analysis; Time-to-progression; Progression-free-survival; Oncology

1. INTRODUCTION

Over recent years, the pharmaceutical industry has become increasingly aware of the need to improve efficiency in the drug development process, through innovative clinical trial design, increased data sharing and focus on personalised healthcare. One important factor in this process is the choice of clinical trial primary endpoint, upon which direct evidence of clinical benefit is required. Within oncology diseases for example, this choice of endpoint has commonly been overall survival, being objective, reliable and easy to measure. However, demonstrating a clinical benefit in survival is becoming increasingly complex due to increasing survival times of patients, higher trial costs, increased availability of alternative therapies and public demand for quicker treatment availability. As such, many researchers are proposing to substitute long term clinical endpoints with shorter term surrogate endpoints that can be assessed in less time and with less cost. For example, a measure of tumour shrinkage, or a composite endpoint of disease progression and death, have often been used as substitutes for overall survival in the assessment of oncology treatments. Use of these endpoints allows treatments to be developed faster, and subsequently made more affordable for payers. This approach has seen increasing popularity, with many recent drug approvals based on so-called surrogate endpoints [1].

In order to replace a long-term clinical trial endpoint with one or more surrogates, it is necessary to evaluate whether the unobserved clinical benefit of treatment on the established longer-term endpoint can be reliably predicted by the observed treatment benefit on the surrogate endpoint(s). Due to the potential variation in treatment benefit amongst different diseases, patient populations and disease-modifying mechanisms of new treatments, this evaluation must be conducted for each potential application of a surrogate endpoint. In many cases, access to data may be limited to a very small subset of comparable data, such as that collected during a single clinical development programme.

Over the last 25 years there have been many contributions to the statistical literature with regard to methodology for evaluating surrogate endpoints. These include single-trial hypothesis testing methods [2], approximation methods [3-7], as well as meta-analytic methods combining data from multiple trials or subgroups within trials [8-17]; a useful summary can be found in the review article written by Weir and Walley [18], along with an updated version written by Ensor et al. [19]. In recent applications (as seen in [20], [21]), the two-stage meta-analytic copula method of Burzykowski et al. [12], an extension to the original two-stage meta-analytic method proposed by Buyse et al. for continuous endpoints [10], has frequently been used. Based on a meta-analysis of many clinical trial datasets, this approach proposes surrogacy measures based on modelling the joint survival distribution of the two (surrogate and long-term clinical) endpoints.

In the case of time-to-event surrogate and true clinical endpoints, investigation into the performance of this method has thus far been restricted to surrogate endpoints that have one outcome of interest, such as exploration of time-to-progression (TTP)

as a surrogate for overall survival (OS) in oncology studies. In reality, in order to maximise the number of events, decrease clinical trial durations and improve the clinical relevance of endpoints, alternative endpoints that consider multiple events of interest are commonly used to assess the clinical benefit of new therapies. Such endpoints, including progression-free survival (PFS), may also incorporate information from both a shorter-term and the true clinical endpoint. PFS is a commonly used endpoint in oncology studies and has been used as the basis for regulatory approval in a number of disease areas.

An alternative surrogacy evaluation approach has been proposed for endpoints that capture multiple events of interest, such as PFS, through the use of a semi-competing risks framework [22]. However, this method is based on separation of the surrogate endpoint into the individual events of interest, and resulting surrogacy evaluations may then not reflect how the commonly defined surrogate endpoint would behave when used in a new clinical study. Whilst the separation of events may offer benefit in some settings, this is not considered a suitable approach when assessing surrogate endpoints that have strong clinical and regulatory understanding and acceptance as measures of clinical benefit, such as PFS in oncology settings.

In this paper, a simulation study is used to assess the performance of the two-stage meta-analytic copula method in the evaluation of two commonly used time-to-event endpoints (time-to-progression and progression-free survival) as surrogates for overall survival in the specific example of oncology clinical trials, for the case where there are limited data available on which to base surrogacy decisions. The aim is to reflect the use of the method from a pharmaceutical industry perspective, where there exist data from a limited number of small-sized clinical trials only, and it is desirable to determine whether a short-term surrogate endpoint can be used in subsequent confirmatory trials. Although the endpoints here are examples of those in oncology clinical trials, the investigation is applicable to any setting where a potential surrogate endpoint also captures data relevant to the true clinical endpoint. The performance of the method has been assessed previously through simulation studies [23], including for small sample sizes [24], however these studies have focused on the scenario where the surrogate endpoint is defined as the time to one particular event of interest, independent of the true clinical endpoint. The impact of using a surrogate endpoint that is defined as the time to either a short-term event or the true clinical event of interest will therefore be assessed here.

Section 2 contains brief details of the surrogacy method under exploration in this study and Section 3 describes the set-up of the simulations, including two different underlying data structures, the two different surrogate endpoints and various combinations of other factors of interest. Results can be found in Section 4, and Section 5 discusses the findings and makes recommendations for future use of the method.

2. Two-Stage Meta-Analytic Copula Model

In order to thoroughly assess a potential surrogate endpoint, Burzykowski et al. [25] recommend to explore two levels of prediction; the ability to predict the unobserved treatment effect on the established long-term endpoint given the observed treatment effect on the surrogate (trial-level surrogacy), and the ability of the surrogate to predict the actual outcome for a given patient, after adjusting for the treatment assignment (individual-level surrogacy). It is desirable for a surrogate endpoint to perform well at both of these levels, in order to provide confidence in its use as a substitute endpoint in further clinical development.

The two-stage meta-analytic copula method proposed by Burzykowski et al. [12] assesses both levels of surrogacy through parameters of the joint survival distribution of the surrogate and long-term (true) endpoints. Using a copula model, specification of the joint survival distribution is achieved using the marginal survival functions of each variable, together with a function which relates the underlying dependence between them. Surrogacy is evaluated through a two-stage procedure, where stage one fits the copula to the data in order to obtain maximum likelihood estimates of treatment effects within each trial, as well as the level of dependence between the endpoints, from which an individual-level measure of surrogacy is derived. Stage two uses random effects modelling to calculate the coefficient of determination between the estimates of the treatment effects, and this is used as the trial-level measure of surrogacy.

Suppose there exist data from $i = 1, \ldots, N$ trials each containing $j = 1, \ldots, n$ subjects with surrogate and true endpoint outcomes S_{ij} and T_{ij} respectively, for patient j in trial i. Then, the general form of the joint survival function of the two endpoints is defined as

$$S(s,t) = P\left(S_{ij} \geq s, T_{ij} \geq t\right) = C_\theta \left\{S_{S_{ij}}(s), S_{T_{ij}}(t)\right\},$$

with $s, t \geq 0, \theta > 1$, where $S_{S_{ij}}$ and $S_{T_{ij}}$ are the marginal survival functions of the surrogate and true endpoints respectively and C_θ is a bivariate distribution function on $[0,1]^2$ with uniform margins. This distribution function is based on a copula function, describing the strength of association between the two endpoints through the parameter θ. For some copula functions, θ can be directly interpreted as an association measure, whereas for other copula models it can be transformed to another measure, such as Kendall's τ [26], to ease interpretability and allow comparison between models. As such, Kendall's τ is the chosen estimator of individual-level surrogacy for the proposed two-stage meta-analytic copula surrogacy method. There are various options for choice of copula function [23], one of which is the Clayton copula, a one-parameter function chosen for simplicity. Based on this copula, the joint survival function is defined as

$$C_\theta\left(S_{S_{ij}}(s), S_{T_{ij}}(t)\right) = \left(S_{S_{ij}}(s)^{1-\theta} + S_{T_{ij}}(t)^{1-\theta} - 1\right)^{\frac{1}{1-\theta}}, \quad \theta > 1. \tag{1}$$

Marginal survival functions for S and T, $S_{S_{ij}}(s)$ and $S_{T_{ij}}(t)$, are assumed to follow proportional hazards models with baseline hazards parametrically specified using a Weibull distribution, although these baseline hazards could also be left unspecified [23]. With this copula function, Kendalls' τ can be conveniently estimated using $\tau = \frac{\theta-1}{\theta+1}$.

Once stage one of the procedure is applied and estimated trial-specific treatment effects on surrogate and true endpoints, (α_i, β_i) respectively, are available, the second stage of the evaluation process can be performed by assuming a reduced random-effects model for these treatment effects:

$$\begin{pmatrix}\alpha_i \\ \beta_i\end{pmatrix} = \begin{pmatrix}\alpha \\ \beta\end{pmatrix} + \begin{pmatrix}a_i \\ b_i\end{pmatrix},$$

where (α, β) are fixed treatment effects, and the random effects (a_i, b_i) are assumed to follow a zero-mean normal distribution with variance-covariance matrix

$$D = \begin{pmatrix}d_{aa} & d_{ab} \\ d_{ab} & d_{bb}\end{pmatrix}.$$

The trial-level measure of surrogacy is then estimated as

$$R^2_{trial} = \frac{d^2_{ab}}{d_{aa} d_{bb}}. \tag{2}$$

A value of R^2_{trial} close to one would suggest that almost all of the variability in the treatment effect on the true endpoint is explained by the treatment effect on the surrogate, whereas a value close to zero would suggest that knowledge of the treatment effect on the surrogate explains little of the variation in the treatment effect on the true endpoint.

Burzykowski et al. [12] discuss bias introduced into the trial-level R^2 in equation (2), caused by the estimation error of the treatment effects coming from stage one of the model. In order to reduce this bias, the method proposed by van Houwelingen et al. [27] is suggested to provide an adjusted version of the trial-level surrogacy measure. However, it is noted that these adjusted estimators are often not available due to non-convergence and inadmissible estimates (outside of [0,1]), which therefore precludes their use in practice [23]. Although alternative approaches have been proposed [28], these adjusted measures are not further explored in our study as they are limited to estimation of R^2_{trial} only and it is our intention to assess both individual

and trial-level surrogacy in a consistent framework. The application of the two-stage meta-analytic copula method in this study is performed making use of publicly available code [29].

A positive feature of the two-stage meta-analytic copula method is that it can be based on any choice of copula function, and indeed Burzykowski et al. [12] describe the importance of selecting an appropriate copula based on the goodness of fit, suggesting a number of ways that this can be done. To explore how the choice of copula can impact interpretation of results, we consider two scenarios in our study. First, we consider performance of the surrogacy method under ideal conditions, where there is no model misspecification and the data are generated to have the same dependence structure assumed by the model. Further to this, we assess the reliability of results when there is model misspecification, by generating data using a different copula function with different underlying data structure to the model being applied.

Renfro et al. [30] also explore the impact of different dependence structures, assessing performance of the two-stage meta-analytic copula method when the underlying data are generated using a Clayton copula constructed using cumulative distribution functions (CDF) instead of survival functions. These two functional constructs allow the same copula function to reflect different dependence structures, thereby assessing the performance of the method in the presence of misspecified dependence. Our work differs from this concurrent work in that we maintain use of the survival implementation of the copula function and assess how results are affected when the surrogate endpoint includes information directly reflecting the true clinical endpoint. We also assume considerably smaller sample sizes, and explore the impact of medium-high censoring across all scenarios.

2.1. Motivating Example

In order to see how the two-stage meta-analytic copula surrogacy method can be applied in practice, we have used it to assess surrogacy within the context of a Phase III study of Herceptin plus chemotherapy versus chemotherapy alone in the treatment of HER2 positive advanced gastric cancer [31]. The primary analysis of this study included 584 patients who were randomly assigned to receive one of two study treatments. The primary endpoint of the study was overall survival, with PFS included as a secondary endpoint. An interim analysis of OS was performed after 75% of the required events had been observed, and at this time the treatment difference (hazard ratio 0.74, 95% confidence interval [CI] 0.60-0.91, median OS of 13.8 versus 11.1 months in the experimental and control arms respectively) was sufficient to cross the pre-specified stopping boundary. The PFS result was consistent with that of OS, demonstrating evidence of a statistically significant benefit from treatment with experimental therapy compared to control therapy (median PFS 6.7 months versus 5.5 months, hazard ratio 0.71 [95% CI 0.59-0.85]).

In practice, data from multiple studies would be available to assess surrogacy and each study would represent an individual unit for analysis. However, in this example of a single clinical trial, the data are grouped according to country, with each country considered to represent a sub-study within the trial. Further discussion of this approach can be found in Renfro et al. [24]. Countries containing seven or fewer patients were grouped by geographical region to allow for parameter estimation; two countries were removed from analysis due to small numbers and the absence of a geographically similar country to combine with (n=4 and n=6 patients respectively). Based on the remaining dataset of 574 patients, results from the application of the surrogacy method show that the R^2_{trial} point estimate (0.57) likely does not support the use of PFS as a surrogate, whereas the individual-level surrogacy ($\tau = 0.67$) could be considered worthy of further investigation.

3. SIMULATION STUDY

As mentioned above, the two-stage meta-analytic copula method has previously been assessed via a simulation study [23]. However, this study was limited in that the impact of the underlying data-generation procedure was not considered, only one type of surrogate endpoint with one event of interest was used, and it was based on sample sizes that are not always realistic in practice. Additional studies designed to address some of these concerns have been conducted [24,30], however none have explored the impact on the joint modelling of using a surrogate endpoint that includes the true endpoint as an event of interest.

The study presented in this paper addresses these concerns by exploring a comprehensive range of factors, as outlined in Table 1.

Table 1: Simulation Scenarios

Factor	Scenarios
Number of trials	4,6
Patients per trial	80, 120, Mixed (50% each of 80,120)
Surrogate Endpoint	TTP, PFS
Data Generation	Clayton, Gumbel
Trial-level association	0.2, 0.5, 0.8
Individual-level association	0.2, 0.5, 0.8
Censoring Rate (on T)	0%, 30%, 60%

There are a number of aims of our study; the first is to determine how well the method performs when using a surrogate endpoint that combines multiple events of interest, including the event of interest for the true endpoint. In the original simulation study performed by Burzykowski et al. [12], the simulated data are constructed according to a time-to-progression scenario, where the surrogate is censored by occurrence of the true endpoint, rather than being considered an event. Our study

generates data according to both time-to-progression (TTP) and progression-free survival (PFS) algorithms, to determine whether there is any impact of using a surrogate that also includes information relating to the true endpoint. In this setting, PFS is defined as the time to the earliest of disease progression or death. This is considered highly relevant since many of the applications of this surrogacy evaluation approach have been based on the use of composite endpoints such as PFS, yet the method has not been explored for this setting via simulation.

The second aim of our study is to assess the performance of the method when there are a very small number of trials with very few patients. Although small-sample simulation studies were performed by Burzykowski et al. [12], the authors considered 10 or 20 trials containing 50, 100 or 200 subjects, which may be considered too many trials compared to those available within a single clinical development plan. Further studies of the two-stage meta-analytic copula method have explored small sample sizes [24], however these studies did not examine in detail the impact of censoring or changes in the underlying trial and individual-level surrogacy. Our study therefore considers 4-6 clinical trials containing 80-120 subjects each, estimating both τ and R^2_{trial}.

One of the most important factors in setting up this simulation study is ensuring that the individual and trial-level association can be accurately controlled. In order to achieve this, Burzykowski et al. [12] control individual-level association through use of a copula model for data generation, with a chosen copula dependence parameter reflecting the strength of surrogacy. Using the copula parameter allows for clear and simple controlling of the individual-level dependence between endpoints, however, since our application of the two-stage meta-analytic copula method is based on the Clayton copula model, our study uses the Clayton as well as the Gumbel copula functions for data generation in order to assess the impact of model misspecification. These two copula functions assume different underlying dependence structures of the two endpoints, and are discussed further in Sections 3.1 and 3.2. In all cases, we construct the joint survival function using exponential survival functions as the marginal distributions of the two endpoints. Inclusion of both of these data generation methods allows us to investigate how the two-stage meta-analytic copula method performs both under ideal conditions, and under model misspecification.

Finally, the original simulation study investigating the two-stage meta-analytic copula method considered just 500 repetitions of the generated datasets, likely due to computational restrictions. Given the extensive list of parameters of interest in our study, which is summarised in Table 1, and the expected computation time, it was felt that the largest number of runs that could be achieved in a reasonable time-frame was 5,000 per scenario. Simulations were run on a Windows 7 64-bit machine with 4GB RAM, using macros based on SAS ® software, Version 9 for Windows [32].

As can be seen in Table 1, in addition to factors described above relating to the number and size of trials and type of endpoint, values of low (0.2), medium (0.5) and

high (0.8) individual and trial-level surrogacy are considered, under varying proportions of censoring. Very few studies have considered low levels of association between endpoints, and those that have were either limited in the number of scenarios under detailed investigation [24] or were based on much larger sample sizes [30]. Additionally, although the range of treatment effects within trials is not of primary interest in this study, previous studies have shown variations in performance of the copula model under various ranges of effects, and so this was added as a final simulation parameter. Simulation parameters were chosen to reflect data characteristics similar to the motivating example.

3.1. Clayton Copula Data Generation

The Clayton copula function with marginal survival functions takes the specific form of equation (1), and to be consistent with Burzykowski et al. [23], the marginal survival functions are chosen to follow an exponential survival distribution. As described by Burzykowski et al. [23], trial-specific random effects are used to control the trial-level association, and in order to obtain draws of S_{ij} and T_{ij} from the joint survival function according to the Clayton copula, the conditional distribution method was applied [23, 33]. The 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 copula function, with strength of dependence controlled using the copula dependence parameter. 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 selected exponential marginal survivor functions. Based on these marginal functions, the joint survival function provides strong upper-tail dependence and weaker lower-tail dependence (see [23] for details).

The baseline hazards 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), therefore providing benefit in terms of the length of the study, and being consistent with the motivating example. The 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), in order to reflect the potential influence of post-progression therapies and long-term follow-up. 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 (0%, 30% and 60%). Since our true endpoint is overall survival, 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.

Recall that although the copula parameter is used to control the level of dependence between the endpoints, it is not always interpretable as a measure of association. Therefore, Kendall's τ is used to select the required individual association between endpoints. For the Clayton copula, θ can be calculated directly from Kendall's τ using $\theta = \frac{1+\tau}{1-\tau}$, and so values of θ were set to 1.5, 3 and 9 in order to achieve `true' individual-level association of 0.2, 0.5 and 0.8 respectively. In order to achieve the required `true' trial-level association values of 0.2, 0.5 and 0.8, the covariance values of the trial-specific random effects were fixed as in [23].

3.2. Gumbel Copula Data Generation

Previous simulation studies of the two-stage meta-analytic copula method use the same copula function to both simulate data and assess surrogacy. In order to investigate whether this can lead to a favourable bias in performance of the copula method, this paper also presents results from simulations where data are generated according to the Gumbel copula. In particular, this approach helps to investigate whether the choice of copula family being applied to the data impacts this method of assessing surrogacy. Based on the joint survival function, the dependency structure of the Gumbel copula is different to the Clayton copula in that it exhibits strong lower-tail dependence (i.e. earlier event times), whereas the Clayton exhibits strong upper-tail dependence (i.e. later event times). For the two endpoints, S and T, the form of the Gumbel model is

$$C_\theta\left(S_{S_{ij}}(s), S_{T_{ij}}(t)\right) = \exp\left[-\left\{\left(-\log S_{S_{ij}}(s)\right)^{\frac{1}{\theta}} + \left(-\log S_{T_{ij}}(t)\right)^{\frac{1}{\theta}}\right\}^{\theta}\right] \quad (3)$$

for $0 < \theta < 1$, where $S_{S_{ij}}(s)$ and $S_{T_{ij}}(t)$ again represent exponential marginal survivor functions for S and T, respectively. The conditional distribution method used to generate data from the Clayton copula cannot be so easily used to generate from the Gumbel copula since the first derivative of the Gumbel copula is not invertible, however the R copula package contains a function to generate correlated random variables according to the Gumbel copula. Since our simulation study makes use of available macros based on SAS software to conduct copula modelling, our data were instead generated using the mixtures of powers algorithm described by Trivedi and Zimmer [34]. Testing of both data generation methods provided datasets with comparable characteristics. The first step of the algorithm is to generate a random variable, γ, from a positive stable distribution, as well as two uniform variables from $U(0,1)$, U_{ij} and V_{ij}. These uniform variables are transformed using γ to be distributed according to the Gumbel copula, with the required individual-level association.

In order to generate γ, a uniform random variable η was drawn from $U(0,\pi)$, and together with the required association level θ, this draw was used to generate a value z according to

$$z = \frac{\sin\big(\eta(1-\theta)\big)\big(\sin(\eta\theta)\big)^{\frac{\theta}{1-\theta}}}{\sin(\eta)^{\frac{1}{1-\theta}}},$$

which was then used to derive γ using a random variable, ω, drawn from a standard exponential distribution, as $\gamma = \left(\frac{z}{\omega}\right)^{\frac{1-\theta}{\theta}}$.

Using this value of γ, U_{ij} and V_{ij} are transformed to be uniform variables which are distributed according to the Gumbel copula, using

$$\tilde{S}_{ij}^{0} = \exp\left(-\left(\frac{-\log\big(U_{ij}\big)}{\gamma}\right)^{\theta}\right),$$

$$\tilde{T}_{ij}^{0} = \exp\left(-\left(\frac{-\log\big(V_{ij}\big)}{\gamma}\right)^{\theta}\right).$$

These two uniform random variables then have the required shape and strength of dependence of the Gumbel copula, and the joint survival function can be constructed by further transforming $\tilde{S}_{ij}^{0}$ and $\tilde{T}_{ij}^{0}$ to time-to-event draws, S_{ij} and T_{ij}, using marginal exponential survivor functions. Censoring was applied as described above. As with the Clayton copula, the required trial-level association is controlled within the covariance matrix D used in the marginal survivor functions, setting ρ equal to the square-root of the required association level. Here, the copula parameter θ can be calculated directly from Kendall's τ using $\theta = 1 - \tau$, so values of θ were set to $0.8, 0.5$ and 0.2 in order to achieve `true' individual-level association of $0.2, 0.5$ and 0.8 respectively.

3.3. Choice of Simulation Parameters

To ensure the most realistic representation of true clinical trial data, certain scenarios were implemented within the data generation algorithm. Firstly, to reflect the impact of long-term follow-up of patients, in particular with respect to the requirement for extended monitoring of disease progression, it was assumed that approximately 5% of subjects would be censored for the surrogate (TTP or PFS) earlier than their time of death. For the composite endpoint of progression and death (PFS), this means that the death event was not used for these 5% patients, which is considered a realistic representation of cases where there is no reliable estimate for the true time of disease progression, for example when there are multiple consecutive missing disease assessments, or if alternative therapy has been started prior to evidence of disease progression.

For cases where OS was censored and the generated value of the surrogate was lower than OS, the surrogate was considered as an event 80% of the time. This allows approximately 20% of subjects to be censored for the surrogate earlier than

the time of censoring of OS, representing scenarios where subjects withdraw consent from further medical procedures to determine disease status, or have disease assessments scheduled less frequently than other clinical trial visits. These factors are considered to reflect true clinical trial settings.

4. Results

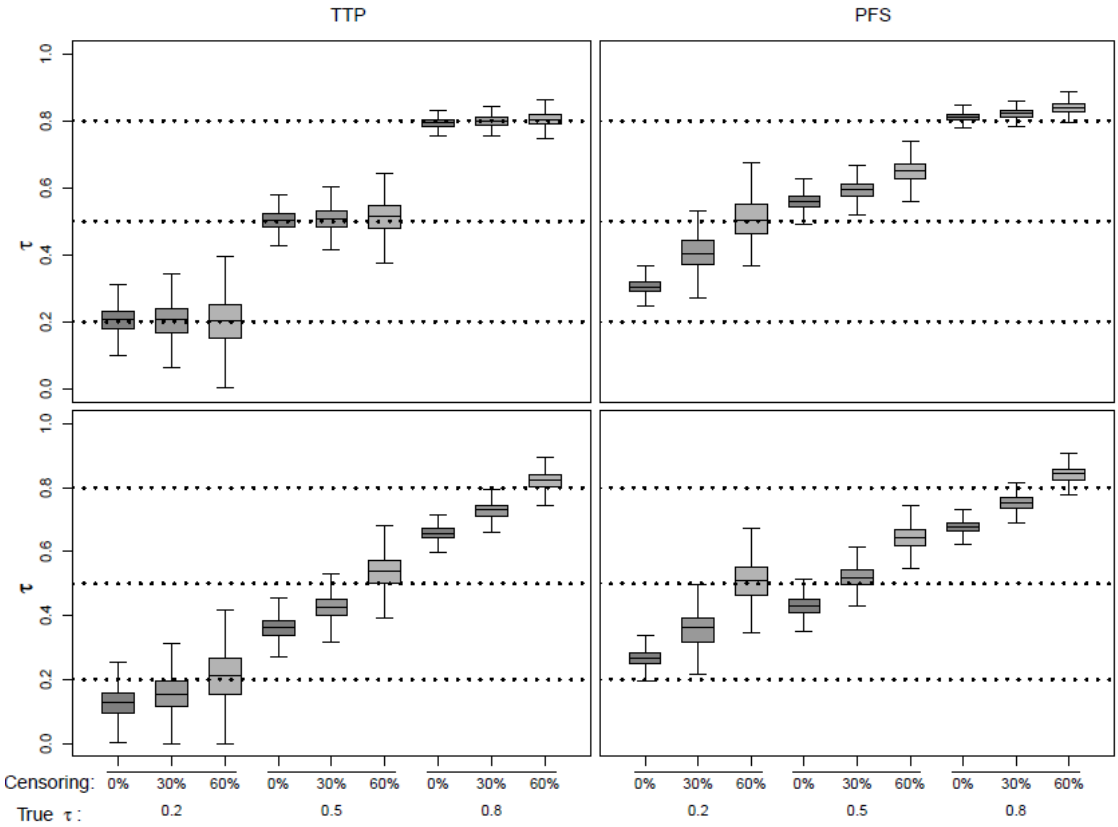
4.1. Convergence

When using TTP as the surrogate endpoint, there were very few issues with convergence of the two-stage meta-analytic copula method, with a maximum non-convergence rate of 1.12%, the majority of which occurred for low levels of true individual-level association. However, when PFS was used as the surrogate, non-convergence was significantly worse, reaching as high as 61.3% for low individual association. In both cases, the non-convergence for medium-high levels of individual association was close to zero, and the issues were mainly found with the low level of true individual association, and this was consistent between the Clayton and Gumbel generated data. The results in this section are therefore based only on those runs that successfully converged, and those that did not converge were not replaced. Since there are approximately 2000 successful runs for even the worst cases of non-convergence, it was felt that this was substantial enough to assess the performance of the method, recalling that previous simulation study to assess the copula used only 500 runs. On occasion there was also a lack of convergence caused by the choice of initial values. Following Burzykowski [23], when this occurred the result from the previous repetition was used, and a sensitivity analysis of available results showed that this was a reasonable approach, with no noticeable differences in the overall conclusion.

4.2. Individual-Level Performance

Figure 1 illustrates the estimated τ values across the simulation scenarios of interest. Each boxplot shows the range of estimated values across all runs, with the level of censoring along the x-axis and the true underlying individual level association on the y-axis. Within the figure, the individual plots display results from the two-stage meta-analytic copula method with Clayton data generation on the top row and Gumbel data generation on the bottom row, with TTP in the left column and PFS in the right column. Since there was little difference in varying the number of trials or sample size within trials, only the smallest sample sizes are presented to illustrate the worst-case scenario (four trials of eighty patients). Results of larger sample sizes can be found in the Supplementary Material. Additionally, since there was little variation in results with varying true underlying trial-level association, the results presented here represent only scenarios with $R^2_{trial} = 0.5$. Results for varying values of R^2_{trial} can also be found in Supplementary Material. Horizontal dashed lines at $y = 0.2, 0.5, 0.8$

represent the true individual-level surrogacy being estimated by each set of three boxplots from left to right.



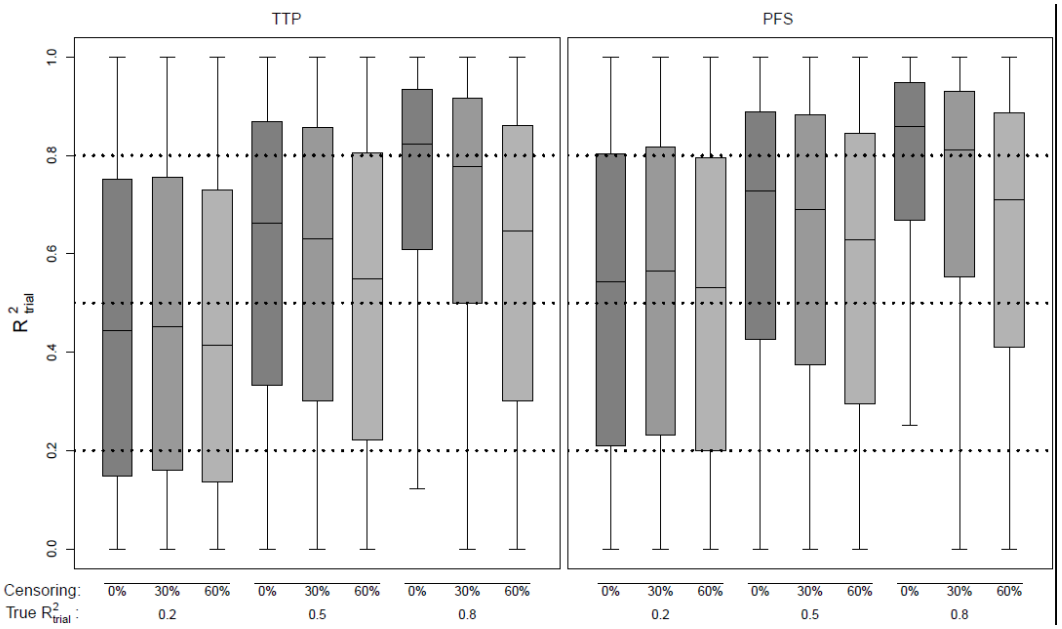
As can be seen, the method performs reasonably well for the TTP scenarios using Clayton-generated data (Figure 1, top left). Consistent with the original simulation study of this method by Burzykowski [23], results were mostly estimated with low average relative bias (maximum 2.8%) despite the small sample sizes explored here, with median estimates lying directly on the respective reference lines. However, variability is relatively high for low-medium levels of association, particularly when there is a high level of censoring. Under the Gumbel data generation (Figure 1, bottom left), it is clear that the performance for TTP deteriorates, with slightly increased variability and a noticeable under-estimation under the presence of little to no censoring. Overall the maximum average relative bias is -38.1%, demonstrating that the method most often under-estimates the true level of association, and could therefore be interpreted as a slightly conservative estimate. However, this interpretation could be hampered by the increased variability. Reassuringly, true high levels of association are estimated with the lowest variability, providing confidence that a large estimated value does in fact correspond to high true association between endpoints.

Whilst results for TTP appear reasonably robust and similar to previous studies, the change to use of PFS as the potential surrogate causes significant issues, even for the Clayton data generation which should reflect the most ideal scenario. In addition to the aforementioned convergence, there is substantial impact on the performance

of the method in estimation of low to medium levels of individual-level surrogacy. Whilst good estimation of truly high association remains, in the little explored scenario of low levels of true association, the estimated τ could be as high as 0.7 for both data generation methods, which could lead to a false conclusion that PFS is predictive of overall survival. The large variability for the true low levels of association also leads to overlap between low and medium association levels, particularly under increased censoring, which hampers interpretation of estimates that lie within a medium-high range (0.4 – 0.7). For estimates even towards the upper limit of this range, it is not realistically possible to conclude that the true underlying association is higher than 0.2. This issue is exacerbated by increased censoring, and there was no improvement from testing with larger sample sizes. Interestingly, the issues introduced through inclusion of PFS as the surrogate have impacted both data generation methods in a similar way, although slightly more impact is seen for the Gumbel data than for the Clayton copula, as could be expected.

4.3. Trial-level Performance

Figure 2 contains similar boxplots to those for individual-level surrogacy, with the y-axis now representing true underlying trial-level surrogacy. As before, only results for the smallest sample sizes are presented (four trials with eighty patients), and the individual-level surrogacy is held at $\tau = 0.5$. Since results were extremely similar between the two data generation methods, only results from the Clayton-generated data are presented here.



When considering the ability to predict the treatment effect on the true endpoint given the observed treatment effect on the surrogate, it is evident that given the small sample sizes considered here, the method cannot be deemed appropriate for

use in this setting. For both endpoints and both data generation methods, the surrogacy evaluation method performs poorly. Although the average estimated value is sometimes close to the true association level, and there is a slight trend upwards as the true underlying association increases, it is also quite often the case that the true association is over or under-estimated. Additionally, there is a large amount of variability in the results, with R^2_{trial} estimates lying across the entire unit interval. Finally, there appeared to be a slight dependence between the individual and trial level association, with increasing R^2_{trial} estimates with increased true individual association. In order to verify results of previous simulation studies carried out by Burzykowski et al. [12], additional simulations were run for larger samples containing 20 trials of 500 patients. The results of these simulations suggested that estimation of R^2_{trial} could indeed be much improved through inclusion of a larger number of studies with larger sample sizes, if those data are available. In summary, the method did not allow for clear data interpretation of R^2_{trial} and cannot be recommended for trial-level analysis of meta-analyses of the size investigated here. The use of study centres within studies as units for surrogacy evaluation has been investigated [24], and will be discussed further in Section 5 in the context of the scenarios explored in this study.

5. DISCUSSION

The main aim of this simulation study was to assess the performance of the two-stage meta-analytic copula method with respect to use of a surrogate endpoint that combines information from a short-term and true clinical endpoint. In addition, it was of interest to evaluate estimation of trial and individual-level surrogacy for small samples, a scenario that is commonly faced by individual pharmaceutical companies wishing to increase efficiency in clinical development programmes through the use of surrogate endpoints. A large range of scenarios were considered, including varied sample sizes, varying strength of individual and trial-level surrogacy, and different levels of censoring.

In line with the simulation study performed by Burzykowski [23], the two-stage meta-analytic copula method performed well in estimating τ for the time-to-progression endpoint, with the level of variability reflecting the small sample sizes used in this study. The change in underlying data structure led to slight under-estimation, but overall the estimates were not alarmingly different to the true values, although variability was considerably high in some cases. At worst, the estimates could be considered as lower bounds of the true association.

For diseases where a high proportion of patients will die before they experience disease progression, TTP is not considered a feasible choice of surrogate endpoint. In oncology drug development, for example, PFS is used much more commonly, since events accumulate faster, trials can be conducted in a shorter period of time

and subjects who die without disease progression are not lost through censoring. The two-stage meta-analytic copula method is currently recommended for use with any time-to-event surrogate [23], but results from this study show that caution is required when considering endpoints that incorporate information from the true clinical endpoint (e.g. PFS) as a possible surrogate endpoint, since a true low (0.2) level of individual association has been shown to be estimated as high as 0.7 in our simulations. This would undoubtedly be convincing enough for a clinician to consider moving forward with use of the surrogate, which could lead to a poor Phase III design and ultimately results that do not support the benefit of the treatment under development. This over-estimation was observed even for the ideal case where there was no model misspecification. For this reason, the two-stage meta-analytic copula method cannot be considered suitable for assessing surrogacy of PFS from clinical trials of the size used in our study. That said, since PFS is defined as the earliest of disease progression and death, it acts as a composite of TTP and OS, and so an encouraging assessment of TTP as a surrogate endpoint could warrant further clinical development based on a PFS endpoint. We would therefore recommend this approach over an assessment of PFS alone for oncology studies. Other diseases areas, such as cardiovascular disease, may also use endpoints that combine multiple events of interest, and the findings from this study may therefore be applicable to these settings also.

With reference to the case study presented in Section 2.1, the results of the simulations hamper the interpretation of the reasonably high estimate of τ, as it is not possible to know whether the estimate reflects a truly high underlying association between endpoints, or over-estimation of low association. This illustrates the uncertainty in conclusions that can be drawn from the two-stage meta-analytic copula method when using PFS as the surrogate, particularly when aiming to evaluate surrogacy from small samples.

Of course, in practice it is necessary to fully understand the underlying structure of the data before selecting a particular copula model to apply; Burzykowski et al. [12] provide details of the surrogacy method for a selection of different copula functions, and suggest that the choice of final model should be based on the one with best fit to the data. Results of our simulations, together with the work conducted by Renfro et al. [30], substantiate the need for careful selection of both the copula family and the dependence structure, showing by two different approaches that when the dependence structure of the data is different to that assumed by the model, results cannot be considered reliable. Importantly, results from our study demonstrate that even under the ideal conditions, where the same survival copula function is used to generate and analyse the data, performance of the method in evaluating PFS as a surrogate endpoint is suboptimal and potentially misleading.

Burzykowski et al. [12] note that one limitation of the copula model is that surrogate and true endpoints are treated symmetrically, so that either endpoint can be shorter or longer than the other. This is clearly not the case when considering overall

survival as the true endpoint, and so the authors highlight that caution is recommended when interpreting results. However, it would appear from our study that there are additional complications with the joint modelling of PFS and OS which need to be explored further. The work of Renfro et al. [30] suggests that alternative modelling using a two-stage, rather than simultaneous, estimation procedure may improve the performance of the two-stage meta-analytic copula method. However, this improvement was not seen uniformly across all simulation settings and so further examination of this is needed to determine whether it can improve the current performance in the assessment of PFS as a surrogate for OS. A further option would be to consider an alternative method to model the joint distribution of the two endpoints, for example through use of a multi-state model [35]. As discussed previously, a semi-competing risks paradigm that accounts for the restriction of S being shorter than or the same as T has also been proposed [22], however this method separates the surrogate endpoint into the individual components. The suitability of this approach therefore depends on the clinical setting and the intended definition of the surrogate endpoint when used in subsequent confirmatory clinical studies.

Importantly, it has been shown that with the limited numbers of trials explored in our study, the method cannot be considered appropriate for assessing the level to which the treatment effect on the surrogate can predict the unobserved treatment effect on the overall clinical endpoint (R^2_{trial}). From the pharmaceutical industry perspective, this suggests that when using this surrogacy assessment method, data from a limited number of small phase I-II clinical trials would generally not provide enough evidence to warrant use of the surrogate endpoint as a complete replacement of the true clinical endpoint in confirmatory phase III trials. To improve estimation, if there exist additional phase III data from similar indications, these could also be included in the surrogacy assessment, accepting the assumptions of generalisability of the treatment, doses, patient population and general study design characteristics. Our exploratory simulations of larger sample sizes suggested that inclusion of additional data could improve performance of the method, however it remains uncertain as to what could be considered a sufficient sample size, and unfortunately a large amount of data are not frequently available.

Further to this, there are often discussions as to whether centres within trials could be used to maximise the number of data points for analysis when only a small number of trials are available. This approach has been studied for both continuous [25] and time-to-event endpoints [24]. Renfro et al. [24] make a recommendation that for time-to-event studies with a moderate (5-9) number of trials, analysis of R^2_{trial} should be conducted using both trial and centre as the units of analysis, with the measure based on trials being considered the primary measure for interpretation. The results of our study indicate that when there are available data from six trials, a measure of R^2_{trial} based on trials as units does not provide reliable conclusions. Additionally, even when there are only four trials available for analysis, the value of

R^2_{trial} based on trials as units is considered key when making inferences about the true underlying strength of surrogacy [24], but based on the context explored in our study this would be very unreliable. Finally, it is currently unclear whether analysis of surrogacy conducted for centres within trials would be considered appropriate by regulatory authorities.

In summary, when applied to small sample sizes, the two-stage meta-analytic copula method proposed for the evaluation of time-to-event surrogates demonstrated poor performance in the assessment of PFS as a surrogate endpoint, but has shown encouraging results when assessing the ability of TTP to predict OS. We therefore recommend that when the desired surrogate endpoint is TTP, an assessment of individual-level surrogacy of time-to-progression is performed using this method. As noted by Burzykowski et al. [23] and Renfro et al. [30], exploration of different copula functions and dependence structures should be conducted, with the choice of final copula function being based on the best fit to the data under investigation. As has been demonstrated in our study with the Gumbel-generated data, the application of a copula model with different functional form to the available data can lead to suboptimal estimation. When PFS is the desired surrogate endpoint, the two-stage meta-analytic copula method must be used with caution, as it may lead to false conclusions that a short-term endpoint has value as a surrogate. Given similarities between TTP and PFS endpoints, we recommend that when PFS is of interest as a potential surrogate, a surrogacy evaluation of TTP is also conducted to determine whether results are consistent between the two.

At the trial-level, a formal quantitative assessment using the two-stage meta-analytic copula method cannot be considered reliable for such a small number of trials (4-6). Less formally, treatment effects that appear consistent between endpoints across multiple trials may be considered as encouraging, however the question remains as to how strong this relationship needs to be before the surrogate can be accepted as a new standard endpoint in future trials.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the associate editor and referees whose comments have substantially improved the quality and content of the manuscript.

References

1. Johnson JR, Ning YM, Farrell A, Justice R, Keegan P, Pazdur R. Accelerated Approval of Oncology Products: The Food and Drug Administration Experience. *Journal of the National Cancer Institute 2011*; **103**:1–9.

2. Prentice R. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine 1989*; **8**:431–440.

3. Freedman LS, Graubard BI, Schatzkin A. Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine 1992*; **11**:167–178.

4. Lin DY, Fleming TR, DeGruttola V. Estimating the proportion of treatment effect explained by a surrogate marker. *Statistics in Medicine 1997*; **16**:1515–1527.

5. Cowles MK. Bayesian estimation of the proportion of treatment effect captured by a surrogate marker. *Statistics in Medicine 2002*; **21**:811–834.

6. Wang Y, Taylor JMG. A measure of the proportion of treatment effect explained by a surrogate marker. *Biometrics 2002*; **58**:803–812.

7. Li Z, Meredith MP, Hoseyni MS. A method to assess the proportion of treatment effect explained by a surrogate endpoint. *Statistics in Medicine 2001*; **20**:3175–3188.

8. Daniels MJ, Hughes MD. Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine 1997*; **16**:1965–1982.

9. Buyse M, Molenberghs G. Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics 1998;* **54**:1014–1029.

10. 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.

11. Gail MH, Pfeiffer R, van Houwelingen HC, Carroll RJ. On meta-analytic assessment of surrogate outcomes. *Biostatistics 2000;* **1**:231–246.

12. 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.

13. 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.

14. 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.

15. Burzykowski T, Buyse M. Surrogate threshold effect: An alternative measure for meta-analytic surrogate endpoint validation. *Pharmaceutical Statistics 2006;* **5**:173–186.

16. Alonso A, Molenberghs G. Surrogate Marker Evaluation from an Information Theory Perspective. *Biometrics 2007*; **63**:180–186.

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. Weir CJ, Walley RJ. Statistical evaluation of biomarkers as surrogate endpoints: a literature review. *Statistics in Medicine 2006*; **25**:183–203.

19. Ensor, H, Lee, R, Sudlow, C, Weir, C.J. Statistical approaches for evaluating surrogate outcomes in clinical trials: a systematic review. *Journal of Biopharmaceutical Statistics 2015*; **26 (5)**: 859-879.

20. Buyse M, Michiels S, Squifflet P, Lucchesi KJ, Hellstrand K, Brune ML, Castaigne S, Rowe JM. Leukemia-free survival as a surrogate endpoint for overall survival in the evaluation of maintenance therapy for patients with acute myeloid leukemia in complete remission. *Haematologica 2011* ; **96**:1106–1112.

21. Laporte S, Squifflet P, Baroux N, Fossella F, Georgoulias V, Pujol JL, Douillard JY, Kudoh S, Pignon JP, Quinaux E, Buyse M. Prediction of survival benefits from progression-free survival benefits in advanced non-small-cell lung cancer: evidence from a meta-analysis of 2334 patients from 5 randomised trials. BMJ Open 2013; **3:**e001802. DOI: 10.1136/bmjopen-2012-001802.

22. Ghosh D, Taylor JMG, Sargent DJ. Meta-analysis for Surrogacy: Accelerated Failure Time Models and Semicompeting Risks Modeling. *Biometrics 2012*; **68**: 226-232.

23. Burzykowski T. Validation of Surrogate Endpoints From Multiple Randomized Clinical Trials With a Failure Time True Endpoint. Unpublished Ph.D. dissertation 2001; available at Limburgs Universitair Centrum, https://ibiostat.be/publications (accessed 6th February 2016).

24. Renfro LA, Shi Q, Xue Y, Li J, Shang H, Sargent DJ. Center-within-trial versus trial-level evaluation of surrogate endpoints. *Computational Statistics and Data Analysis 2014*; **78**: 1-20.

25. Burzykowski T, Molenberghs G, Buyse M. The Evaluation of Surrogate Endpoints. Springer: New York, 2005.

26. Kendall M. A new measure of rank correlation. Biometrika 1938; **30 (1-2)**: 81-93.

27. van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine 2002;* **21**:589–624.

28. Renfro L, Shi Q, Sargent D, Carlin B. Bayesian adjusted R^2_{trial} for the meta-analytic evaluation of surrogate time-to-event endpoints in clinical trials. *Statistics in Medicine 2012;* **31**:743–761.

29. Validation of a Failure-time Surrogate for a Failure-time True Endpoint. https://ibiostat.be/software/surrogate (accessed 6th February 2016).

30. Renfro LA, Shang H, Sargent DJ. Impact of copula directional specification on multi-trial evaluation of surrogate endpoints. *Journal of Biopharmaceutical Statistics 2015*; **25**: 857-877.

31. Bang YJ, Van Cutsem E, Feyereislova A, Chung HC, Shen L, Sawaki A, Lordick F, Ohtsu A, Omuro Y, Satoh T, Aprile G, Kulikov E, Hill J, Lehle M, Rschoff J, Kang YK. Trastuzumab in combination with chemotherapy versus chemotherapy alone for treatment of HER2-positive advanced gastric or gastro-oesophageal junction cancer (ToGA): a phase 3, open-label, randomised controlled trial. *The Lancet 2010*; **376**:687–697.

32. Copyright, SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.

33. Nelsen RG. An introduction to copulas. Springer Verlag, 1999.

34. Trivedi PK, Zimmer DM. Copula modeling: An introduction for practitioners. *Foundations and Trends in Econometrics 2005*; **1**:1–111.

35. DeJardin D, Lesaffre E, Verbeke G. Joint modeling of progression-free survival and death in advanced cancer clinical trials. *Statistics in Medicine 2010*; **29**:1724–1734.