Statistical Methodology for Evaluation of Time-to-Event Surrogate and True Endpoints in Small-Sample Meta-Analysis of Clinical Trials

Natalie Dimier

Submitted for the degree of

Doctor of Philosophy

Department of Mathematics and Statistics

University of Reading

September 2017

Abstract

Clinical trials can be lengthy and costly, with new treatments taking more than a decade to become available to the patients who need them. It is therefore of great interest to improve efficiency in this process, such as replacing the primary endpoint of a clinical trial with an alternative endpoint that can be measured with greater ease, reduced cost or reduced observation periods. Such replacement endpoints are called surrogate endpoints, and there has been a vast amount of research conducted to establish statistical methodology that can reliably assess whether such endpoints are appropriate for future use.

The aims of this research are therefore threefold; to identify appropriate methodology that can be used in the assessment of time-to-event surrogate and true endpoints; to examine the identified methods via simulation studies for the setting of small sample sizes, across a variety of scenarios, and in particular for surrogate endpoints that capture information on both an intermediate disease status and the long-term clinical outcome of interest; and finally to develop improved methodology that can advance the surrogacy evaluation process for these settings.

The findings of the research build on the existing surrogate endpoint literature by demonstrating that the most commonly used approaches for evaluation of time-to-event surrogate and true endpoints can have potential limitations. As a result of this finding, and based on the identified strengths and weaknesses of the examined statistical approaches under the settings of interest, a novel methodology for the evaluation of time-to-event surrogate and true endpoints is proposed and evaluated. This method provides an alternative option for the evaluation of surrogate endpoints, and is recommended for further use.

Acknowledgements

This PhD would not have been possible without the unwavering support of my supervisor Professor Sue Todd, who provided exceptional guidance and encouragement throughout. Sue offered limitless patience to my questions and my often erratic working schedule, and provided outstanding support and laughter over many years and for which I will be forever grateful.

I would also like to thank my current employer, Roche Products Ltd, for providing funding, and my previous employer GlaxoSmithKline plc who provided funding and time for this research to be conducted.

Finally, I would like to thank my family and friends who have supported me over the years, through the challenges and successes, and without whom this thesis would not have been completed.

Declaration: I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Natalie Dimier

Contents

List of Tables

List of Figures

Chapter 1

Background and Context for Surrogate Endpoints

1.1 Introduction

The development of new medicinal products is a long and complex process that requires significant investment of both time and money. Experimental treatments must undergo intensive testing through multiple phases of clinical trials, to ensure that they demonstrate therapeutic benefit to patients alongside an acceptable safety profile. For a new treatment to achieve regulatory approval, there has to be clear confirmatory evidence of clinical benefit, based on a measure that is reliable, objective and relevant for patients. As more treatment options become available and new standards of care are introduced, satisfying the need for substantive efficacy results from adequate and well-controlled clinical trials is becoming increasingly difficult.

The long-established process of testing new molecules through multiple phases of clinical trials can be lengthy, and in order for continued development of new medicines to remain feasible, researchers are exploring ways in which the efficiency of this process can be improved. A key determinant in the length, cost and complexity of any clinical trial is the selection of primary endpoint; the measure of clinical benefit that ultimately deter-

mines the 'success' or 'failure' of the study. Careful selection of this parameter is critical in ensuring that the study is interpretable, relevant to patients and medical practitioners, and considered approvable by regulatory and health authorities.

In many therapeutic areas, primary endpoints that are considered the gold standard for satisfying these criteria are becoming increasingly difficult to use, due to increased costs and long follow-up of patients. With this in mind, researchers are investigating the plausibility of substituting such long-term primary endpoints for shorter-term endpoints that can be evaluated more readily. For such short-term, or 'surrogate' endpoints to be accepted by health authorities as substitutes for the traditionally used clinical endpoint, they need to undergo a rigorous assessment, both clinically and statistically, to establish their reliability in predicting long-term outcome and treatment effects thereon. Such evaluations are heavily dependent on there being sufficient data available from previous clinical studies to quantify the accuracy of these predictions.

The research presented in this thesis explores the motivation for use of surrogate endpoints, with particular focus on the statistical methodology that is designed for their evaluation. Of primary interest is if and how surrogate endpoints can be reliably examined from the perspective of an individual pharmaceutical company, who may only have a limited number of small clinical trial datasets available for the particular disease or molecule under investigation. To explore why this is important, the remainder of this chapter highlights an introduction to the concept and motivation for surrogate endpoints, including definitions (Section 1.2), potential benefits and limitations (1.3), consideration of when surrogate endpoints may be useful (1.4), regulatory aspects (1.5) and motivation for further research (1.6).

1.2 Defining a Surrogate Endpoint

There have been a number of attempts to define a surrogate endpoint, both conceptually and statistically. According to the US Food and Drug Administration (FDA) guidance

on expedited programs for drug development, "a surrogate endpoint is a marker, such as a laboratory measurement, radiographic image, physical sign, or other measure, that is thought to predict clinical benefit, but is not itself a measure of clinical benefit" (FDA, 2014). Similarly, in the European Medicines Agency (EMA) notes for guidance on ICH topic E8, a surrogate endpoint is defined to be "an endpoint that is intended to relate to a clinically important outcome but does not in itself measure a clinical benefit" (EMEA, 1998).

These two definitions are very similar, with a key requirement that the surrogate must relate to, or predict, the true clinical benefit as measured by the original and accepted clinical endpoint. A number of researchers consider correlation between endpoints to be a good indicator that the surrogate can be used in place of the long-term clinical endpoint, however, correlation alone is not considered sufficient for a surrogate endpoint to be considered worthy of future use (Fleming and DeMets, 1996). An equally fundamental criterion is that, since the surrogate endpoint is being used to replace the true clinical outcome of interest, the observed treatment effect on the surrogate outcome must be a reliable predictor of the unobserved treatment effect on the clinical outcome. Both of these criteria must therefore be established before a surrogate endpoint can be used as a primary endpoint in confirmatory clinical trials.

To make steps towards these goals, Prentice (1989) was the first to propose a formal statistical definition of surrogacy, as "a response variable for which a test of the null hypothesis of no relationship to the treatment groups under comparison is also a valid test of the corresponding null hypothesis based on the true endpoint". This definition requires a 'valid' surrogate to be an endpoint that captures all of the treatment benefit on the true clinical outcome, such that knowledge of the treatment effect on the surrogate outcome would provide full knowledge of the treatment effect on the clinical outcome and remove the need to formally test the null hypothesis for the long-term endpoint. This has been the cornerstone in the development of statistical methodology to evaluate potential surrogates, which is described in detail in Chapter 2.

Some examples of potential surrogate endpoints include CD4 cell count as a surrogate for the development of AIDS in HIV (De Gruttola et al., 1993), blood pressure as a surrogate for cardiovascular disease, and tumour shrinkage as a surrogate for survival in cancer studies.

1.3 Benefits and Limitations of Surrogate Endpoints

The potential practical benefits of substituting a final clinical endpoint with a shorter term surrogate endpoint can include shorter trial durations, higher compliance to study protocols, reduced costs, increased ethics, trial feasibility and simplified recording and monitoring of clinical trial data (Burzykowski et al., 2005). As investigational agents are being developed, and new treatment options are emerging, all of these benefits sound very attractive to the pharmaceutical industry, payers and patients alike. For the companies expending the resources, the shorter trial durations can lead to a lower rate of non-compliance, patient drop-out and loss to follow up, and subsequently a higher level of reliability in the data. The lower costs associated with these shorter trials mean that drugs can be tested, approved and marketed at an increased rate, despite the hurdle of needing to prove superiority or non-inferiority over consistently improving standards of care. For payers, the lower research and development costs can lead to more competitive pricing and increased cost-effectiveness. For those patients with the disease, the chance to receive a potentially life-saving treatment as early as possible is of the utmost importance. If drugs can be developed quicker, they have more chance to be made affordable, reimbursable and therefore accessible to those who need them. A number of successful surrogate evaluations have been conducted based on large meta-analyses, with some leading to FDA approval of surrogate measures in HIV (FDA, 2015; Marschner et al., 1998), and further approvals of a number of treatments for adjuvant breast and colon cancers based on an endpoint of disease-free survival (FDA, 2007), for example.

Alongside these potential benefits, the use of surrogate endpoints has been subject

to criticism (Fleming and DeMets, 1996). Applications of surrogate endpoints have not always been successful, and in some cases have led to harmful results. One of the most striking examples in practice is that of the CAST trial (Pratt and Moye, 1990; Greene et al., 1992). Following FDA approval of three investigational agents (encainide, flecainide and moricizine) for the treatment of ventricular arrhythmia, it was hypothesised that longer-term treatment with these agents would reduce the incidence of sudden cardiac death after myocardial infarction. The biological plausibility of the relationship between these endpoints seemed reasonable; since arrhythmia leads to increased risk of death, suppression of arrhythmia should lead to reduced mortality.

The CAST phase III trial was designed to test this hypothesis by comparing mortality rates between the three active agents and a placebo. However, the results of an interim analysis of this study showed that the rate of death in the active treatment arms was 2.5 times higher than that in the placebo arm (Pratt and Moye, 1990). Both the encainide and flecainide arms were discontinued early when 33 sudden deaths occurred in patients taking these active compounds, compared to 9 in the placebo group. At the final analysis, the comparison was 63 deaths in the active arms versus 16 in the placebo group (Fleming and DeMets, 1996). Analysis of the moricizine data showed that this too increased the risk of death (Cardiac Arrhythmia Suppression Trial II Investigators, 1992). This highlights the difficulty that researchers and clinicians can face when selection of the primary endpoint depends heavily on understanding of the disease under investigation, and the potential damage that can be caused if surrogate endpoints are not sufficiently investigated and validated. In this case, what appeared to be clear rationale for a surrogate measure of benefit turned out to be incorrect, and led to harmful conclusions.

1.4 When might Surrogate Endpoints be suitable?

Phase II trials, designed to investigate dose and gain an indication of efficacy and safety, generally use short-term primary endpoints that can provide relevant clinical information

regarding the activity of an experimental treatment. Since these trials are rarely used alone to provide confirmatory evidence of clinical benefit, the selection of clinical endpoint can be flexible, with many short-term measures chosen in order to expedite further clinical development of a molecule. In most cases, short-term endpoints are used without a formal assessment of surrogacy (FDA, 2007; Fleming and DeMets, 1996).

In contrast, the confirmatory phase III trials which follow need to be designed to provide definitive evidence of treatment benefit based on a clinically relevant and reliable measure. In many cases, such measures can be difficult to assess, or require a very long follow-up period to be observed, which can lead to patient dropout or non-compliance, and subsequently have an adverse effect on the overall trial conclusions. Furthermore, there are some outcomes that must be measured using burdensome and invasive medical procedures, and so it would be beneficial for patients if alternative ways to measure treatment benefit could be made available. In each of these settings, having an opportunity to replace the traditionally used endpoint with a surrogate has appeal to researchers, medical practitioners and patients (Fleming and Powers, 2012).

One frequently used long-term endpoint is overall survival (OS), used in many therapeutic areas to assess clinical benefit of new treatments. The benefits of this endpoint are clear; it is unambiguous, reliable, objective, and positive results provide confirmatory evidence of extended life of patients. However, there are also many drawbacks. Firstly, it must be considered whether it is truly ethical to wait for the occurrence of a fatal event in order to estimate the efficacy of a new treatment. Secondly, it can be an endpoint which requires a long follow-up time, with clinical trials extending for many years before enough evidence can be collected to confirm efficacy benefits. Finally, there is the possibility that the endpoint can be confounded by factors such as next-line treatments, since those patients who do not respond to their randomised therapy may require further treatment with other marketed or experimental agents. This can make the results of overall survival endpoints difficult to interpret, and in some cases may lead to underestimation of the treatment benefit.

One example of this can be seen in the BIG 1-98 phase III trial (BIG 1-98 Collaborative Group, 2005), designed to compare the efficacy of letrozole and tamoxifen as monotherapy and sequential treatments in post-menopausal women with estrogen receptor-positive breast cancer. A total of 8010 patients were enrolled in the trial, and initial results comparing monotherapy letrozole to monotherapy tamoxifen concluded a significant increase in disease-free survival for the letrozole monotherapy arm, leading to 25% of patients from the tamoxifen arm to also be treated with letrozole. Updated intent-to-treat (ITT) analyses conducted four years later confirmed the benefit in disease-free survival, but also concluded that there was no significant difference in overall survival between the two monotherapy groups (BIG 1-98 Collaborative Group, 2009). However, sensitivity analyses accounting for the crossover treatment from tamoxifen to letrozole concluded that OS was indeed significantly extended for those randomised to the letrozole arm (Colleoni et al., 2011). This case-study demonstrates how the results of OS, despite being objective and clinically relevant, can be confounded and potentially lead to erroneous conclusions.

The greatest case for the use of surrogate endpoints is therefore in confirmatory Phase III studies, where the time required to complete studies renders them infeasible, or possibly irrelevant by the time results are reported. Even in these settings, the use of surrogates is not simple, as there is currently no standardised approach for their biological or statistical evaluation, nor any regulatory recommendation as to which statistical methodology is considered appropriate, or what results would be considered sufficient.

Furthermore, in order for a surrogate endpoint to be considered appropriate for use, there are many aspects that require careful consideration (Fleming and Powers, 2012). Firstly, there must be a good understanding of the underlying pathway of the disease under investigation, sufficient to have confidence that therapeutic benefit in the surrogate endpoint will also lead to clinical benefit for the patient in the long-term outcome of interest (Fleming and DeMets, 1996). Secondly, the expected magnitude of benefit on the true endpoint must be understood; regulators need to know that approved medicines will show sufficient benefit in long-term outcome to ensure a balanced benefit/risk ratio

for the patient. Finally, health technology assessors need to understand the value of the treatment in order to make informed decisions on reimbursement and healthcare policy. All of these factors require that potential surrogate endpoints are thoroughly assessed prior to implementation in clinical trials, and such assessments require statistical evaluation and analysis to determine the reliability and accuracy of surrogate endpoints (Burzykowski et al., 2005).

1.5 Regulatory Aspects

A key component in the evaluation of surrogate endpoints is ensuring that they are considered appropriate for approval by regulatory agencies. Both the FDA and EMA recognise the need for more efficient development and approval of new medicines, and have implemented procedures for researchers to do so using surrogate endpoints.

In the United States, the FDA grant 'regular' approval of new medicines when clinical benefit has been demonstrated, or when a treatment effect has been demonstrated on an 'established' surrogate. A surrogate endpoint is considered established when there exist substantial data that have increased certainty that the surrogate is truly predictive of clinical benefit. When such data are not available, a surrogate is considered 'unestablished' and cannot be used for full regulatory approval.

In recognition of the need for use of 'unestablished' surrogates, the FDA introduced an *Accelerated Approval* program for diseases thought to be 'serious' or 'life-threatening' (FDA, 2014). According to this regulation, interventions tested using adequate and well-controlled clinical trials based on a surrogate endpoint which is "reasonably likely to predict clinical benefit" can be sufficient for approval and marketing, with the condition that post-marketing clinical trials are conducted to confirm the clinical activity of the drug. This enables patients to gain access to the treatment while it is still under investigation. Since introduction of this regulatory pathway in 1992, a total of 169 accelerated approvals have been granted by the FDA (as of 30th June 2017 (FDA, 2017)).

The EMA offer a similar program; that of conditional approval. However, they already allow more flexibility in their recommended choice of endpoint. They recommend that the primary endpoint be chosen to provide a valid and reliable measure of clinical benefit, stating that, for example, progression-free survival (PFS) or disease-free survival (DFS) are acceptable primary endpoints in certain situations, with overall survival reported as a secondary endpoint. In these cases, where the trial is powered for the primary comparison, the sample size and study duration should allow for survival data to be precise enough to rule out negative effects on OS. Further to this, for diseases where the time from progression to death can be long, and the effect of treatment on the primary endpoint is large, there may in fact be no need to show evidence of superiority in overall survival. However, when there are no treatment options available as next-line therapies, when the time from disease progression to death is expected to be short, or when it is expected that there will be significant differences in toxicity in favour of the control arm, it is recommended that overall survival be chosen as the primary endpoint (EMEA, 2013).

Interest in finding and implementing surrogate endpoints has increased rapidly over recent years, with medical practitioners, clinicians and statisticians providing valuable contributions to the debate. For some, the use of surrogate endpoints is seen as a potential risk to the drug development process, caused by a lack of understanding of disease pathways and difficulty surrounding the validation of such endpoints (Fleming and DeMets, 1996). For others, surrogate endpoints are seen not only as a convenience, but a necessity, to ensure that the pharmaceutical industry is able to continue to provide safe and effective treatments to those who need them in the shortest possible time.

1.6 Motivation for Further Research

Researchers must show caution when selecting potentially useful surrogates, by considering both the disease setting and the class of treatment under investigation. A surrogate that is encouraging in one indication may not necessarily extend to treatment with the same

compound in other, even similar, disease settings (Fleming and DeMets, 1996). Similarly, the class of treatment under investigation may well influence the relationship between the proposed surrogate and clinical endpoint; a surrogate which is predictive of treatment effect for one intervention may have no relevance for the same surrogate and clinical endpoints for a different intervention targeting the same disease. Finally, as new treatments become available and treatment practices evolve, the need to evaluate surrogate endpoints will be a continuous process. What was effective as a surrogate endpoint for one disease setting may well change as treatments are approved and introduced into the wider patient population, and medical knowledge increases. This means that there will also always be the potential to improve on the statistical methodology used to evaluate the surrogates.

There has been a vast amount of research conducted into appropriate statistical methodology for the evaluation of surrogacy; this can be seen by the large number of methods discussed in Chapter 2. The approaches have developed based on hypothesis testing and estimation methods, in single trials and in a meta-analytic setting, however there are many questions that remain unanswered.

There is currently no consensus, in the statistical literature or in regulatory guidance, as to which of the methods are considered most appropriate for any given setting. Whilst most recommend meta-analytic methods over single trial analysis (Burzykowski et al., 2005; Daniels and Hughes, 1997), there are few recommendations as to which of the meta-analytic measures perform most satisfactorily, and which may be subject to bias under particular conditions. In addition, the majority of proposed measures provide quantification of surrogacy based on a $[0, 1]$ scale, but there is currently no recommendation for how large such measures need to be before a surrogate can be considered suitable for use, and it is suggested that clinical and other judgement is required to decide this (Molenberghs et al., 2008; Weir and Walley, 2006). Without such guidance, it is difficult for researchers to understand the regulatory requirements for surrogates, and there is a risk of subjective and inconsistent conclusions from different applications of the same methodology.

Further, and considered of particular interest for this research, is the performance of

measures when there exist very little data on which to base a surrogacy evaluation. The majority of previous investigations of surrogacy have been based on large meta-analyses, conducted by independent groups who gathered data from many sources, for example the work of Shi et al. (2016) and Shi et al. (2017). Much of the motivation for the future use of surrogates comes from individual pharmaceutical companies, who are planning Phase III trials and are trying to determine which endpoint is most appropriate. In such cases, there may exist data only from within the same or a similar clinical development plan, which may consist of only a handful of small Phase II and III trials. Obtaining vast amounts of data from competitors, although improving, remains challenging, and so it is critical that research is conducted to determine the reliability of surrogacy evaluation methods in this setting. Whilst there has been some exploration of this setting already (Renfro et al., 2014), such investigation has been limited to just one statistical approach to evaluating surrogacy, and there is a need to expand the scope to determine which of the variety of available methods can be considered most appropriate for use in practice.

Finally, although surrogate and long-term endpoints can be based on any type of outcome, such as binary, continuous or time-to-event (Burzykowski et al., 2005), this research focuses on statistical methods applicable to time-to-event surrogate and true endpoints, as these are the endpoints most likely to suffer from the need for long periods of observation. The results and recommendations presented herein are applicable to all time-to-event endpoints, however oncology endpoints are selected to illustrate the performance of the methods. Two scenarios are of particular interest, in which overall survival is used as the long-term outcome of interest. The first scenario considers a surrogate of time-to-progression (TTP: time from study entry to disease progression) and the second considers progression-free survival (PFS: time from study entry to disease progression or death). The motivation for considering both of these cases is that in the first, the event of death has no impact on the surrogate other than to censor the outcome if patients die prior to disease progression, whereas in the second, death is also included as an event for the surrogate outcome. Therefore, comparison of both endpoints allows an assessment

of the impact of including information directly related to the long-term outcome within the surrogate endpoint. This has so far been explored very little in the literature, but is considered hugely important since PFS is used much more commonly in oncology trials, with the larger number of potential events leading to the endpoint being quicker to achieve maturity (see additional discussion of Ghosh et al. (2012), for example). PFS is also well understood and accepted by clinicians and regulatory authorities as a secondary, and in some cases primary, measure of clinical benefit.

A detailed description of the development of statistical methodology for the evaluation of surrogate endpoints is presented in Chapter 2. Based on this review of the literature, two methods have been selected for further investigation. The first of these methods is examined in Chapter 3, where the performance of the measure is assessed for the new scenario of time-to-event surrogates that also contain information from the true clinical endpoint. This scenario has not previously been explored for the selected method, but represents one of the most commonly used time-to-event oncology endpoints in practice, progression-free survival. Focus remains on surrogacy assessment from an individual pharmaceutical company perspective, where there are limited data available for analysis. The second measure is examined in Chapter 4, again under the previously unexplored scenario of assessment of surrogates that contain data from the true endpoint. The original proposal for use of this second method was based on simulation studies of single clinical trial datasets; the performance of the method in a meta-analytic setting is therefore investigated.

Based on findings from these investigations, Chapter 5 contains a proposal for a new measure of assessing surrogacy. Whilst applicable to all endpoint types, focus remains on time-to-event endpoints in order to compare results to the previous findings of the research. This methodology is proposed to address some of the issues encountered during investigation of previously recommended measures, and is evaluated through simulation studies. An illustrative example of the application of all three of the investigated surrogacy methods is provided in Section 6, and final conclusions and recommendations are presented in Chapter 7.

Chapter 2

Review of Statistical Methodology Designed for Evaluation of Surrogate Endpoints

2.1 Introduction

The need for a thorough statistical evaluation of potential surrogate endpoints provides motivation for research into the most appropriate and reliable statistical methodology for this purpose. As noted in Section 1.2, demonstrating that there exists correlation between two endpoints, or between treatment effects on two endpoints, is not sufficient to 'validate' a surrogate endpoint for use (Fleming and DeMets, 1996). Of primary interest is whether the outcome of, and treatment effect on, a surrogate endpoint can reliably predict outcome and treatment effect for the long-term clinical endpoint of interest.

Statistical methodology designed to evaluate potential surrogate endpoints has developed over the last 30 years, from simple measures based on individual clinical trials (single-trial measures, Section 2.2), to more complex assessments using meta-analysis of many datasets (meta-analytic measures, Sections 2.3 and 2.4). Research has examined *individual-level surrogacy*; the ability of a surrogate outcome to predict long-term outcome

for an individual patient, and *trial-level surrogacy*; how well the unobserved treatment effect on the long-term endpoint can be predicted using the observed treatment effect on the surrogate endpoint.

In this chapter, the development of statistical methodology designed for the assessment of surrogate endpoints is described, with an aim to provide context for further research and identify gaps that need to be addressed. As discussed in Section 1.6, the focus of interest in this research is time-to-event endpoints, i.e. endpoints that measure time from study entry to the occurrence of some event of interest, such as time to disease progression or death. Measures that have been proposed for surrogacy but which cannot incorporate time-to-event endpoints are therefore described only when critical to the understanding of historic context or subsequent methodology, such as extensions to methods to allow their use with time-to-event endpoints. Thorough reviews of the development of statistical methodology can be found in Weir and Walley (2006) and Ensor et al. (2016). Further guidance on the use of some of these measures is also provided by Alonso et al. (2017).

Throughout, the long-term clinical outcome of interest is referred to as the 'true' endpoint, and the surrogate and true endpoints are denoted by random variables S and T respectively. Treatment effects on S and T are denoted by α and β respectively. When describing meta-analytic measures, it is assumed that there exist a total of N clinical trials available for analysis, with n_i patients enrolled in each trial $i = 1, ..., N$, with S_{ij} and T_{ij} used to denote the surrogate and true outcomes, respectively, for patient j in trial i. It is also assumed that each trial includes a binary treatment indicator, Z_{ij}, which may or may not be the same treatment across all available studies. When discussing general model constructs, $X_j(t)$ will be used to denote the covariate vector for patient j, which may contain the treatment indicator Z_{ij} as well as other relevant covariates. For time-to-event outcomes, it is assumed that patients either experience the event of interest, or are censored, meaning that they did not reach the event of interest during their period of observation. It is this censoring, commonplace in the analysis of time-to-event outcomes, that causes difficulties in the extension of many surrogate evaluation methodologies and

leads to the need for specific measures to assess these types of endpoints. The introduction to existing measures of surrogacy starts in the next section with those based on a single clinical trial.

2.2 Single-Trial Measures

2.2.1 Prentice Paradigm and Proportion of Treatment Effect Explained

As stated in Section 1.2, the first formal statistical definition of a surrogate endpoint was provided by Prentice (1989), who proposed that the treatment comparison based on the surrogate should provide full information on the treatment comparison based on the true endpoint. Burzykowski et al. (2005) write this definition statistically as

$$f(S|Z) = f(S) \Leftrightarrow f(T|Z) = f(T),$$

where $f(.)$ represents the probability distribution for the random variables S and T, conditional or not on treatment, Z. Therefore, a lack of treatment effect on the surrogate outcome implies a lack of treatment effect on the true outcome, and vice versa, such that only a formal hypothesis test of the surrogate endpoint is required to draw conclusions about the treatment effect on the true endpoint. In order to satisfy this overall definition, a number of operational criteria are proposed by Prentice (1989);

$$\begin{aligned} f(S|Z) &\neq f(S), \\ f(T|Z) &\neq f(T), \\ f(T|S) &\neq f(T), \\ f(T|S,Z) &= f(T|S). \end{aligned}$$

These criteria require that, for a surrogate to be considered 'valid', it is necessary to observe statistically significant treatment effects on both S and T, a statistically significant

impact of S on T, and crucially, that the conditional distribution $T|S$ is not statistically significantly different depending on whether treatment is also accounted for. In other words, all of the treatment effect on the true endpoint is mediated by the surrogate endpoint.

Whilst these criteria have the benefit of being applicable to any type of endpoint, they have been heavily criticised, in particular due to the final criterion which requires non-significance of a hypothesis test. Firstly, the lack of non-significance does not confirm that the two distributions are equivalent, since there may simply be insufficient evidence or statistical power to detect a difference (Freedman et al., 1992). Secondly, it could be questioned whether such a requirement is realistic in practice. Whilst the biological plausibility of a surrogate must be strong, it could be possible that the underlying pathways of a disease do not allow for absolutely all of the treatment effect to be captured by the surrogate. The aim of surrogate endpoints is to reliably predict unobserved treatment effects on T, but there must be a balance between accuracy of predictions and the potential benefits of use of the shorter-term endpoint. Some loss of precision may be considered acceptable if it leads to faster drug development and patient access to new medicines. Indeed, as noted by the FDA (FDA, 2014), a surrogate can be considered for use in clinical development if it is "reasonably likely to predict clinical benefit".

Fleming et al. (1994) suggest using this final criterion "as an ideal to keep in mind" rather than definitive evidence of surrogacy, and Buyse and Molenberghs (1998) note that whilst the criteria are informative and will tend to be satisfied for a valid surrogate, strict satisfaction is not necessary. In addition, studies have shown that a perfect surrogate can fail to satisfy the criteria (Tsiatis, 1996). However, despite the criticism, the fundamental concept behind the Prentice (1989) definition remains appealing, and has formed the basis of much of the ensuing statistical literature.

To reduce the strictness of a binary decision resulting from a hypothesis test, and to provide an alternative approach to estimating the value of a surrogate, Freedman et al. (1992) propose a measure estimating the proportion of treatment effect on T that can

be captured by adjustment for S, the proportion explained (PE). If β and β_S denote the treatment effect on T unadjusted and adjusted for S, respectively, the PE is estimated as

$$PE = 1 - \frac{\beta_S}{\beta}.$$

A high value of PE would suggest that the majority of treatment effect on T is captured or mediated through the surrogate. It could then be concluded that knowledge of the treatment effect on S provides sufficient information about the treatment effect on T. A value of PE close to zero would suggest that adjusting for the potential surrogate had little impact on the treatment effect on T, and therefore that the surrogate did not offer much potential.

The PE provides more information than a hypothesis test; it estimates what level of treatment effect can be captured by the surrogate. It also has the benefit of being applicable to any type of endpoint. However, the measure has been subject to criticism, since values can lie outside of the $[0, 1]$ interval, confidence intervals can be very wide, and it becomes difficult to interpret when endpoints are not measured using the same scale, for example ratios versus linear differences. It has been recognised that these criticisms lead to difficulties in interpretation of the measure, and lead to it being infeasible for use in practice (Freedman, 2001). Despite these limitations, many researchers continued to develop and build on the idea of the proportion explained, including Lin et al. (1997), who investigated the PE when based on time-to-event endpoints; Li et al. (2001), who extend using a generalised linear model approach; Chen et al. (2003) who allow time-varying covariates and Cowles (2002) for Bayesian estimation.

2.2.2 Relative Effect and Adjusted Association

Although the limitations of the PE prevent reliable use of the measure in practice, the underlying concept remains appealing, providing estimation of how much of the treatment effect on T can be captured by S. More fundamentally, of particular interest is how well the treatment effect on S can predict the treatment effect on T, such that future use of

a surrogate provides some information as to the benefit of treatment on the long-term clinical outcome. To address this, Buyse and Molenberghs (1998) suggest to replace the PE with two alternative measures, the underlying concepts of which have been key to much of the subsequent methodological developments.

The first of the two measures, the relative effect (RE), is defined as the ratio of treatment effects on T versus S

$$RE = \frac{\beta}{\alpha},$$

providing an estimate of the multiplicative relationship between the two treatment effects. If this assumption of a multiplicative relationship is considered reasonable, an estimated value of RE from a previous study can be used, together with an estimated treatment effect on a surrogate endpoint from a new trial, to provide a prediction of the unobserved treatment effect on the true endpoint for that new trial. The RE is considered a trial-level measure of surrogacy, since it allows prediction of the treatment effect for a new trial.

The second proposed measure, the adjusted association (AA), is intended to evaluate the association between S and T after accounting for treatment, such that for an individual patient it would be possible to predict future clinical outcome using the outcome observed on the surrogate endpoint. This ability to predict long-term outcome for a given individual is denoted the individual-level surrogacy, and for normally distributed endpoints is calculated based on the error terms generated from a linear model of surrogate and true endpoints regressed on treatment.

While the concepts of RE and AA are applicable to all types of endpoints, there are a number of limitations that have prevented their widespread use. In particular, the confidence intervals for the RE, being based on data from only one clinical trial, can be extremely wide, hampering interpretation. In addition, the measure assumes a linear relationship between α and β, which cannot be verified when based on data only from a single trial. As a result, a number of researchers have concluded that evaluation of surrogate endpoints should be based on data from multiple clinical trials, in a meta-

analytic setting (Daniels and Hughes, 1997; Buyse and Molenberghs, 1998). Potential advantages of meta-analytic approaches to surrogacy evaluation also include the ability to assess the robustness and sensitivity of surrogate endpoints within different patient populations and treatments.

The extension of RE and AA to multiple trials has been a key development in the statistical literature for surrogate endpoint evaluation. Whilst not the only contribution to the literature, extensions to multiple endpoint types and combinations thereof has led to a wealth of techniques designed to assess both trial-level and individual-level surrogacy, as described below.

2.3 Meta-Analytic Methods

The aim of statistical methodology for the assessment of surrogate endpoints is to quantify the accuracy and reliability of predictions for unobserved outcomes and unobserved treatment effects. For a new endpoint to be considered appropriate for the widest possible range of future settings, it is of the utmost importance that it is evaluated across as many different clinical trial datasets as possible, to ensure generalisability without over-extrapolation of results and conclusions. The choice of clinical trials to be included in a meta-analytic evaluation of surrogacy is therefore critical.

As an example, the duration of treatment for a disease may be dependent on individual patient characteristics or prognosis, such as age or other co-morbidities. In such cases, if the aim is to use a surrogate endpoint across all future clinical studies of that disease, the surrogacy assessment must incorporate clinical trials that contain all possible treatment regimens in all populations. Without this representation of all potential uses of the treatment, it is difficult to justify universal use of the surrogate. Based on such meta-analysis, use of a surrogate may also be restricted to scenarios that clearly demonstrated reliable surrogacy.

Furthermore, whilst differing biological aspects of individual diseases mean that evalu-

ation of potential surrogates must be done separately for each, there are additional questions to be considered, such as the class of treatment under investigation. An assessment of surrogacy may be conducted for individual treatments, but more likely is that classes of treatment will need to be explored based on their mechanism of action, so that use of surrogate endpoints remains relevant for the future, and so that sufficient data are available on which to base a surrogacy assessment. An evaluation of multiple clinical trials which test different treatments within the same therapeutic class is therefore possible, but care needs to be taken to ensure that treatments within these trials are biologically similar and can be expected to act upon endpoints in a consistent way.

The earliest formal statistical methodology for the evaluation of surrogate endpoints based on data from multiple trials was proposed by Daniels and Hughes (1997), who suggest to model the difference in treatment effects on S and T across multiple trials from a Bayesian perspective. This first step into meta-analytic surrogacy evaluation had the benefit of requiring only summary level information from individual trials, rather than individual patient level data, which can be difficult to access. However, the resulting downside is that individual-level surrogacy cannot be evaluated. Nevertheless, this first move into meta-analytic assessment of surrogacy formed the basis for a number of extensions (Burzykowski et al., 2005). Key developments of further meta-analytic measures relevant to the work in this thesis are described next.

2.3.1 Two-Stage Methods

Extensions of the concepts of RE and AA to meta-analytic surrogacy assessment were used to build a framework designed to evaluate surrogate and true endpoints of continuous data (Buyse et al., 2000). Although not directly applicable to time-to-event endpoints, the method is briefly described here as an introduction to extensions which do allow analysis of such outcomes. Throughout, the meta-analytic version of individual-level surrogacy is denoted R^2_{indiv}, and the meta-analytic version of trial-level surrogacy is denoted R^2_{trial}.

In a meta-analytic setting, each clinical trial provides estimates of treatment effects on S (α) and T (β), which are measured with some level of estimation error and will vary from trial to trial. This variability is key in estimating the strength of surrogacy, as it reflects the quality of predictions that can be made from the available data. In order to define a trial-level measure of surrogacy, which can be used to quantify the variability in predictions of treatment effects in future studies, Buyse et al. (2000) consider estimation of the variability between trial-specific estimates of treatment effects (α_i, β_i), using either fixed effects or random effects modelling.

In the fixed effects model, two-stages of analysis are used to estimate model parameters. In the first stage, it is assumed that patient outcomes, S_{ij} and T_{ij} follow the linear model:

$$\begin{aligned} S_{ij} &= \mu_{S_i} + \alpha_i Z_{ij} + \epsilon_{S_{ij}}, \\ T_{ij} &= \mu_{T_i} + \beta_i Z_{ij} + \epsilon_{T_{ij}}, \end{aligned} \tag{2.1}$$

where μ_{S_i}, μ_{T_i} are trial-specific intercept terms and $\epsilon_{S_{ij}}$, $\epsilon_{T_{ij}}$ are correlated error terms which are assumed to follow a multivariate Normal distribution with mean zero and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_{SS} & \sigma_{ST} \\ & \sigma_{TT} \end{pmatrix}.$$

Using this model from the first stage of analysis, the square of the correlation coefficient between S and T, based on covariance matrix (2.3.1), is used to quantify the association between S and T after accounting for Z (individual-level surrogacy), using

$$R^2_{indiv} = \frac{\sigma^2_{ST}}{\sigma_{SS}\sigma_{TT}}.$$

This value then lies in $[0, 1]$, with values close to one suggesting strong association between endpoints. The obvious limitation to this measure is that there is no objective threshold that would indicate when a surrogate is reliable enough for use, a point that will be discussed further in Section 2.3.4. A confidence interval for R^2_{indiv} can be calculated using the delta method as

$$R^2_{indiv} \pm z_{1-\alpha/2}\sqrt{\frac{4R^2_{indiv}(1-R^2_{indiv})^2}{n_{total}-3}},$$

where $z_{1-\alpha/2}$ is the critical value of the standard normal distribution and n_{total} is the total sum of patients coming from all clinical trials included in the meta-analysis (Alonso et al., 2017).

As a second stage to the analysis, and to define a trial-level measure of surrogacy, the trial-specific terms of Equation (2.1) are considered to follow a linear model

$$\begin{pmatrix} \mu_{S_i} \\ \mu_{T_i} \\ \alpha_i \\ \beta_i \end{pmatrix} = \begin{pmatrix} \mu_S \\ \mu_T \\ \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} m_{S_i} \\ m_{T_i} \\ a_i \\ b_i \end{pmatrix},$$

where the terms m_{S_i}, m_{T_i}, a_i and b_i are assumed to be Normally distributed with mean zero and covariance matrix

$$D = \begin{pmatrix} d_{SS} & d_{ST} & d_{Sa} & d_{Sb} \\ & d_{TT} & d_{Ta} & d_{Tb} \\ & & d_{aa} & d_{ab} \\ & & & d_{bb} \end{pmatrix}.$$

This model can be used to estimate the parameters and respective variability based on previous clinical trial data, which can subsequently be used to estimate the proportion of variability in the treatment effect on T that can be explained using the treatment effect on S, using the coefficient of determination. This is proposed as a trial-level measure of surrogacy, defined as

$$R^2_{trial} = \frac{\begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}}{d_{bb}}.$$

This measure of R^2_{trial} lies in $[0, 1]$, with values close to one suggesting that almost all variability in treatment effects on T can be accounted for using treatment effects on S, which would be a clear indication of good surrogacy. Again, the measure is limited due to the subjectivity surrounding a threshold that would be sufficient to declare a surrogate

valid, which is discussed further in Section 2.3.4. Confidence intervals can be calculated using the delta method, as

$$R^2_{trial} \pm z_{1-\alpha/2}\sqrt{\frac{4R^2_{trial}(1-R^2_{trial})^2}{N-3}},$$

where $z_{1-\alpha/2}$ reflects the critical value of the standard normal distribution, and N reflects the number of trials (Alonso et al., 2017).

Whilst this approach follows a two-stage procedure, continuous endpoints could also be analysed using a random effects modelling approach where the two stages are combined to estimate the model parameters and respective variance in one step. This process leads to the same measure of trial-level surrogacy as above, but is not described further here since extensions to other endpoints do not allow for such a modelling approach, and the method has been subject to computational difficulties; details can be found in Buyse et al. (2000). Further, a reduced version of R^2_{trial} has also been considered, where the model is simplified to remove the trial-specific intercept terms. In such cases, the simplified R^2_{trial} is defined as $R^2_{trial} = \frac{d^2_{ab}}{d_{aa}d_{bb}}$. This model, based on random intercepts independent of random treatment effects is used in the time-to-event setting, which will be described in the next section.

There have been a number of extensions to this two-stage approach, in particular to capture alternative endpoint structures, where the linear modelling approach is no longer appropriate. For binary outcomes, it is argued that there is no clear model choice to handle these endpoints efficiently (Burzykowski et al., 2005), and so the binary outcomes are assumed to derive from underlying latent continuous random variables. This assumption allows the proposed R^2_{indiv} and R^2_{trial} from the continuous endpoint structure to be used, although it is recognised that parameter estimation is more difficult. Such an extension also allows for ordinal endpoints to be used, as well as different combinations of binary, ordinal and continuous outcomes.

For time-to-event endpoints, there is no extension that can allow the previously defined surrogacy measures to be directly used, and so an alternative extension is required. This

additional methodology, of primary relevance to the research in this thesis, is referred to as the two-stage meta-analytic copula method, and is described in detail in the next section.

2.3.2 Time-to-Event Endpoints

Time-to-event endpoints measure the time from the start of observation, such as study entry or date of randomisation of a patient into a clinical trial, until the occurrence of a clinical event of interest. For example, a time-to-event endpoint may measure the time from study entry until death, or capture durations of a particular disease status, for example the duration of time from the first response to treatment until disease starts to deteriorate. Such endpoints can also be referred to as failure-time or survival endpoints.

A key feature of such data is that, during the period of a clinical trial, not all patients will experience the event of interest. There will likely be a number of patients who drop out of the clinical study and withdraw their consent for further observation. There will be other patients who reach the end of the observation period without experiencing the event. Such cases are considered to be 'censored', indicating that they provide data for analysis up to the point that they are no longer observed, but the time of their clinical event remains unobserved and so their 'survival time' remains unknown. It is important to note that the event of interest may not always be death, so the term 'survival' is used to indicate that the particular event of interest has not been observed.

This censoring leads to challenges in the analysis of time-to-event data, which therefore requires alternative statistical methodology that can accommodate censored observations and make use of the data collected up to the point of censoring. A number of these techniques are employed in the statistical evaluation of time-to-event surrogate and/or true endpoints, therefore definitions to illustrate concepts, terminology and notation related to survival analysis are briefly described below. These descriptions are restricted to those that are critical to the understanding of proposed surrogacy evaluation measures.

Definitions

- **Survivor Function**: In the analysis of time-to-event data there are two functions of primary interest, the survivor function and the hazard function. The survivor function, $S(t)$, defines the probability that a patient will survive beyond some time t, $P(T > t)$, where T is a random variable denoting actual survival time. This is frequently estimated using the Kaplan-Meier methodology described below.

- **Hazard Function**: The second function that is of interest in the analysis of time-to-event data, the hazard function $\lambda(t)$, denotes the probability of an individual experiencing an event at time t given that they survived up to time t. The hazard function is used in regression modelling of time-to-event data using the Cox proportional hazards model (see below). As a general result, it can be shown that the hazard and survival functions are conveniently linked. As noted in the description of the survivor function above,

$$S(t) = P(T > t) = \int_t^\infty f(u)du = 1 - F(t),$$

where $f(u)$ denotes the probability density function of the random variable T (denoting the time-to-event), and $F(t) = P(T < t)$ denotes the distribution function of T. Then, the hazard function is defined as

$$\begin{aligned}
\lambda(t) &= \lim_{\delta t \to 0} \left\{ \frac{P(t \leq T < t + \delta t | T \geq t)}{\delta t} \right\} \\
&= \lim_{\delta t \to 0} \left\{ \frac{1}{\delta t} \frac{P(t \leq T < t + \delta t)}{P(T \geq t)} \right\} \\
&= \lim_{\delta t \to 0} \left\{ \frac{F(t + \delta t) - F(t)}{\delta t} \right\} \frac{1}{S(t)}.
\end{aligned}$$

The term in the limit here, shown in brackets, is the definition of the derivative of $F(t)$ with respect to t, which equates to f(t), thereby leaving the hazard function defined as $\lambda(t) = \frac{f(t)}{S(t)}$. Using differentiation with the chain rule, it can further be

shown that

$$\frac{dlog(1-F(t))}{dt} = \frac{-f(t)}{1-F(t)} = \frac{-f(t)}{S(t)} \Rightarrow \frac{-dlogS(t)}{dt} = \frac{f(t)}{S(t)} = \lambda(t),$$

and therefore the general relationship between survivor and hazard functions can be derived as

$$S(t) = \exp\left(-\int_0^t \lambda(y)dy\right).$$

- **Kaplan Meier Method (Kaplan and Meier, 1958):** This approach estimates the survivor function by estimating the reduction in cumulative probability of survival each time an event occurs in the sample. At time zero, all patients are event-free, and the cumulative probability of survival is equal to one. As time progresses, patients start to experience the event of interest, and at each event time t the probability of survival is re-calculated as the probability of survival just prior to time t multiplied by the probability of observing an event at time t (number of events occurring at time t divided by the number of patients remaining under observation and event-free immediately prior to time t [the risk set]). Censored patients remain in the risk-set until their observation discontinues. The time at which probability of survival first reaches below 50% is termed the 'median' survival time and is a key summary measure of survival outcomes.

- **Cox Proportional Hazards Regression (Cox, 1972):** As for any regression model, the proportional hazards model allows for comparison of outcomes between patient groups, after accounting for other covariates (if required). For time-to-event data, the outcome to be modelled is the risk that the event of interest will occur at time t, the hazard function. Covariates are assumed to have a multiplicative effect on the hazard function, and hazard functions are assumed to be proportional between patients with different covariate values. The comparison of hazard functions between groups then provides a 'hazard ratio', a parameter which describes whether patients in one group are more (hazard ratio > 1) or less (< 1) likely to experience

the event of interest over the period of observation. The hazard ratio is a commonly used measure of treatment benefit for time-to-event outcomes. The general form of the Cox model for an individual patient, j, with time-dependent covariates $X_j(t)$ is given by

$$\lambda_j(t|X_j(t)) = \exp(\beta' X_j(t))\lambda_0(t),$$

where $\lambda_j(t|X_j(t))$ denotes the hazard function for patient j, $\beta' X_j(t)$ denotes a linear combination of the vector of covariate values at time t for patient j and $\lambda_0(t)$ denotes the hazard function for a patient with values of zero for all covariates, also known as the baseline hazard. In this representation, the covariate vector $X_j(t)$ is described as a function of time, t, since some applications of this model to surrogacy evaluation apply a time-dependency concept to model the impact of the change in surrogate outcome over time. This will be re-visited when describing the respective surrogacy approaches.

Two-Stage Meta-Analytic Copula Method

The first meta-analytic surrogacy approach that was proposed specifically for the evaluation of time-to-event surrogate and true endpoints is an extension of the previously described two-stage model for continuous endpoints (Burzykowski et al., 2001). As noted, the extension is not trivial, due to the nature of the time-to-event endpoints, and in particular the censoring.

In order to provide measures of individual and trial-level surrogacy for the time-to-event case, Burzykowski et al. (2001) propose to replace the linear model used in stage one of estimation with a function that defines the joint distribution of the surrogate and true endpoints. To achieve this, a copula function is used (Nelsen, 1999), and so this approach is described herein as the two-stage meta-analytic copula method. Copula functions provide a method for constructing the joint distribution of two endpoints, using marginal uniform distributions for the endpoints and defining a dependence structure between them. The

uniform marginal distributions can then be transformed to the required endpoint type, without impacting the strength or structure of relationship between endpoints.

The general definition of a bi-variate copula is a function on $[0,1]$ x $[0,1]$ for which, for standard uniform random variables u $Un(0,1)$ and v $Un(0,1)$,

$$\begin{aligned} C(u,0) &= 0, \\ C(0,v) &= 0, \\ C(u,1) &= u, \\ C(1,v) &= v, \end{aligned}$$

and for all u_1, u_2, v_1 and $v_2 \in [0,1]$ with $u_1 \leq u_2$ and $v_1 \leq v_2$,

$$C(u_2,v_2) - C(u_1,v_2) - C(u_2,v_1) + C(u_1,v_1) \geq 0.$$

There are many different choices of copula function, some which describe the association between endpoints using a single parameter, and others that do so using multiple parameters. A detailed review of copula theory can be found in Nelsen (1999). Burzykowski et al. (2001) consider functions with one dependence parameter, for simplicity, and concentrate on three different copula functions; Clayton, Gumbel and Plackett. Indeed, the importance of the selection of copula has been a topic for much research (Renfro et al., 2015).

Of primary interest in this thesis is the Clayton copula model, since this was the model selected for previous simulation studies to assess performance of the proposed surrogacy evaluation approach (Burzykowski, 2001), and has been the subject of further research since (Renfro et al., 2014, 2015). Further discussion of the Gumbel model can be found in Chapter 3, where it is used to examine the performance of the two-stage meta-analytic copula approach under misspecified models.

For two uniform random variables, u and v, the general form of the Clayton copula is

$$C_{\theta_c}(u,v) = \left(u^{1-\theta_c} + v^{1-\theta_c} - 1\right)^{\frac{1}{1-\theta_c}},$$

where θ_c denotes the dependence parameter specific to the Clayton copula, quantifying the strength of association between u and v. For this copula function, $\theta_c > 1$, with the association between endpoints decreasing as θ_c reduces. To apply such a model to time-to-event data in the context of surrogate endpoints, Burzykowski et al. (2001) propose to transform the uniform margins to survivor functions for S_{ij} and T_{ij}, such that the joint survival distribution between surrogate and true endpoints is defined as

$$S(s,t) = P(S_{ij} \geq s, T_{ij} \geq t) = C_{\theta_c}\{S_{S_{ij}}(s), S_{T_{ij}}(t)\}, \qquad s,t \geq 0,$$

where $S_{S_{ij}}(s)$ and $S_{T_{ij}}(t)$ denote the marginal survival functions for S and T. Using this model, estimation of the effects of treatment on the two endpoints is proposed through the use of proportional hazards models for each endpoint:

$$\begin{aligned} S_{S_{ij}(s)} &= \exp\left\{-\int_0^s \lambda_{S_i}(y)\exp(\alpha_i Z_{ij})dy\right\}, \\ S_{T_{ij}(t)} &= \exp\left\{-\int_0^t \lambda_{T_i}(y)\exp(\beta_i Z_{ij})dy\right\}, \end{aligned}$$

where λ_{S_i} and λ_{T_i} correspond to the trial-specific baseline hazard functions, and α_i and β_i denote the trial-specific treatment effects on the surrogate and true endpoints respectively, for each trial. In this setting, the proportional hazards models are considered separately for each endpoint, and the covariate vector is not considered to be time-dependent. To estimate the parameters of the joint survival distribution, $S(s,t)$, including the dependence between endpoints, θ_c, and the trial-specific treatment effects, α_i and β_i, the baseline hazards are specified parametrically to allow use of maximum likelihood estimation. For both endpoints, baseline hazards are assumed to derive from a Weibull distribution, although alternative approaches can be considered, including leaving the form of the baseline hazards unspecified (Burzykowski et al., 2001). When maximum likelihood estimation is used to estimate the copula model parameters, including the trial-specific treatment effects, a Newton-Raphson procedure can be implemented.

As in the case of continuous endpoints, an individual-level measure of surrogacy is defined as the association between S and T after accounting for treatment, Z. Whereas

the square of the correlation coefficient could be used for continuous data, such a measure is not appropriate in a time-to-event setting since the baseline hazard may be different for each trial. Instead, the copula dependence parameter, measuring the dependence between endpoints after accounting for trial and treatment effects, is considered a good candidate for individual-level surrogacy. However, since each copula model has a different dependence parameter, Burzykowski et al. (2001) propose to transform the copula parameter into an estimate of Kendall's τ (Kendall, 1938), which can then be used to compare results across different copula models for the same dataset. For two random variables u and v, τ can be calculated as

$$\tau = 4 \int_0^1 \int_0^1 C_{\theta_c}(u, v) C_{\theta_c}(du, dv) - 1,$$

serving as a measure of individual-level surrogacy. Of note, when using the Clayton copula function, the dependence parameter θ_c has an additional interpretation relevant to survival analysis. This parameter also represents the ratio of the hazard rate for T conditional on $S = s$ to the hazard rate for T conditional on $S > s$, thereby quantifying the increase in risk of the true outcome (e.g. death) for observing a surrogate outcome (e.g. disease progression) at time s relative to achieving a longer time to surrogate outcome. A higher value of θ_c therefore indicates that the risk of observing the true outcome increases with a shorter time to surrogate outcome, reflecting a stronger association between surrogate and true endpoints.

During this first stage of analysis, the copula model provides estimation of not only the dependence parameter, and therefore the individual-level surrogacy, τ, but also the trial-specific treatment effects for S and T. In a second stage of analysis, as in the continuous setting, these estimated treatment effects are then regressed according to

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

where the final term is assumed to follow a zero-mean normal distribution with variance-

covariance matrix

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

A trial-level measure of surrogacy is then proposed as

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

which is the same measure as that used in the approach for assessing continuous surrogate and true endpoints. Confidence intervals for both τ and R^2_{trial} can be constructed using the delta method (Burzykowski et al., 2005).

Burzykowski et al. (2004) also consider the use of a copula modelling approach when assessing a time-to-event true endpoint and a binary or ordinal surrogate endpoint. Assuming a latent underlying continuous random variable for the surrogate endpoint, for which specific cutpoints define the given categories, the Plackett copula model is proposed to estimate both τ and R^2_{trial} in the same way as for time-to-event surrogate and true endpoints.

2.3.3 Applications

The two-stage meta-analytic copula method has been used in a number of recent surrogacy evaluations, and has become the preferred method for many, particularly in oncology settings. The method is used by Chibaudel et al. (2011) to assess a number of surrogate endpoints to replace OS in clinical trials of advanced colorectal cancer, including PFS and duration of disease control. Laporte et al. (2013) use the approach in evaluating PFS as a surrogate for OS in advanced non small-cell lung cancer, and Foster et al. (2011) and Foster et al. (2015) use the approach to assess PFS as a potential surrogate for OS in extensive stage small-cell lung cancer, with tumour response also being evaluated as a surrogate using the two-stage copula modelling approach developed for the assessment of binary surrogates. Shi et al. (2017) also applied the method to assess a binary surrogate for PFS in the assessment of patients with follicular lymphoma. Some of these applications

split the studies into smaller subgroups to increase the number of data points available for analysis; this important topic will be discussed further in Section 2.5.

Despite these examples of the use of the two-stage meta-analytic copula approach, a number of limitations have also been noted, as will be described in the next section. Limitations that are considered specific to the two-stage meta-analytic copula method are described first. More general limitations of the surrogacy approaches under investigation within this thesis are described in Section 2.4.4.

2.3.4 Limitations of the Two-Stage Method

A practical limitation of the two-stage meta-analytic copula method is that the underlying modelling relies on complex computational procedures that are not available in standard software packages. Whilst Burzykowski et al. (2001) have provided code to apply the methodology (see `http://ibiostat.be/online-resources/online-resources/surrogate`), it remains challenging to implement due to the complex joint modelling and likelihood maximisation procedures that are required. As an attempt to address this, Cortiñas and Burzykowski (2010) explored a number of alternative approaches to estimate R^2_{trial} that would not require the complex joint modelling of endpoints, however none of these approaches were considered to provide reliable results, and so this remains of concern.

A further issue that has been noted by Burzykowski et al. (2001) but not extensively studied in the literature is that of the assumptions associated with modelling using copula functions. In particular, regardless of endpoint choice, copula models assume that the two endpoints used for the marginal distributions are symmetrical, such that either of the endpoints can be shorter or longer than the other. In general, if we wish to model time-to-event surrogate endpoints, it would be most likely in practice that the surrogate would be shorter than the true endpoint, and for some surrogates it is impossible for the true endpoint to be shorter. One particular example is the endpoint of progression-free

survival, defined as the time until disease progression or death. When the true clinical endpoint is time to death (overall survival), it is not possible for progression-free survival to be longer, and so Burzykowski et al. (2001) note that caution is needed in the interpretation of the copula approach in this setting. Whilst simulation studies of the two-stage meta-analytic copula method have been conducted by Burzykowski (2001), Renfro et al. (2014) and Renfro et al. (2015), none of these included an assessment of progression-free survival as a surrogate for overall survival, and violation of the assumption of symmetry between endpoints has therefore not been previously explored. Such exploration is therefore considered of high importance and is conducted in Chapter 3 of this thesis.

The simulation study conducted by Burzykowski (2001) assessed the performance of the two-stage meta-analytic copula method for a number of factors, including varied trial-level and individual-level surrogacy, sample size and proportions of censoring. However, there remain a number of scenarios left unexplored. Firstly, the sample sizes examined consisted of 10 or 20 trials each containing 50-200 patients, representing an ideal situation where there are many datasets available for analysis. When there exist very few datasets available, the reliability of the surrogacy measures could be expected to be lower, and it is important to understand the impact of this. Less data implies that parameters of the models are estimated less precisely, which in turn can lead to imprecise predictions of future treatment effects. Whilst some trade-off between accuracy of predictions and the length and cost of future clinical trials could be warranted, it is important to assess whether such predictions may be misleading. Therefore, exploration of much smaller sample sizes is performed in Chapter 3.

Additionally, the underlying surrogacy at both individual and trial levels were previously restricted to values of 0.5 or greater, which does not provide insight into settings where the true surrogacy is low. Whilst low values would not be of interest to evaluate a surrogate endpoint further, it is important to understand how well the method performs in these circumstances, to ensure that truly poor surrogate endpoints are recognised as such. Whilst some exploration of lower trial-level surrogacy has been conducted elsewhere,

these are settings that are not directly relevant to the work of this thesis due to the large sample sizes being considered (Renfro et al., 2012, 2015).

In order to address some of the limitations mentioned above and to try and improve the two-stage meta-analytic copula method, a variety of additional work has been undertaken, which is described in the next section.

2.3.5 Other Relevant Areas of Investigation

Burzykowski et al. (2001) recognise that when using the two-stage meta-analytic copula method, the estimated treatment effects from stage one of analysis are used directly in stage two of analysis, without taking into consideration that they are subject to measurement error. Subsequent use of these treatment effects to estimate R^2_{trial} may therefore lead to bias, and so the methods of van Houwelingen et al. (2002) (used prior to formal publication) and Fuller (1987) are proposed to provide an adjusted estimate of trial-level surrogacy. Unfortunately, the simulation study of Burzykowski (2001) demonstrated that there were substantial issues with non-convergence of these alternative methods that effectively prevents their use in practice. To investigate further, Renfro et al. (2012) considered Bayesian estimation of the R^2_{trial} measure via simulations, and conclude that this approach could offer improvements in estimation.

Further to this, Shi et al. (2011) examine the performance of the two-stage meta-analytic copula modelling approach in evaluating R^2_{trial}, as compared to the use of simpler measures of correlation coefficients and least squares regression. Based on simulated datasets, separate Cox proportional hazards models were used to estimate the treatment effect for each endpoint in each trial. For comparison, R^2_{trial} values were estimated from these treatment effects using the square of Pearson and Spearman correlation coefficients, the coefficient of determination from weighted regression of treatment effects (weighted by sample size) and the two-stage meta-analytic copula method. Results demonstrated that whilst the squared Spearman's correlation coefficient performed poorly across most

scenarios, the remaining measures performed similarly. Whilst low levels of trial-level surrogacy were explored (0.2), the number of trials considered were $N = 6, 12, 18$ with a total of $n = 500, 1000, 2000$ patients per trial, which provides a very large sample size overall. This study therefore does not address the setting of interest in this thesis, when there exist very little data on which to assess surrogacy.

Smaller sample sizes were considered by Renfro et al. (2014) as part of an extensive investigation into the use of subgroups within clinical trials as the units for analysis rather than entire clinical trials. Due to the relevance of this investigation on the use of surrogacy approaches by individual pharmaceutical companies, the work is described further in Section 2.5.

Finally, Renfro et al. (2015) refer to the two-stage meta-analytic copula method as the 'gold standard' for surrogacy evaluation, highlighting a number of recent applications to case studies. However, caution is recommended when considering use of the approach, to ensure that the specification of the copula model, as well as the direction of dependence between S and T, is appropriate for the dataset being analysed. As has been noted earlier in this section, there is a need to choose both the copula function, for example the Clayton, as well as the marginal distributions of S and T to use with this copula. The description of the two-stage meta-analytic copula method above highlights that marginal *survival* functions are proposed, in order to define the joint survival distribution between S and T. This construct provides a dependence structure of strong association between S and T at later survival times, and weaker association between S and T at earlier survival times. Renfro et al. (2015) consider how an alternative choice of marginal distributions can reverse this relationship, such that early survival times are strongly associated and later survival times are weakly associated. This is achieved by assuming marginal *distribution* functions for S and T, thereby defining the joint distribution function of S and T using the copula. A joint survival function can be derived from this joint distribution function, however it has a different likelihood function to the joint survival distribution derived from the same copula based on marginal survivor distributions. This reverses

the direction of the association between S and T, which can lead to biased estimates of model parameters, including marginal treatment effects and the copula dependence parameter, if the dependence structure assumed by the model does not match that of the observed data. Renfro et al. (2015) therefore recommend that careful selection of not only the shape of association (by the copula model), but also the direction of association (by the marginal distributions), is carefully considered before choosing the final copula implementation. Further detail on the shape and direction of the copula dependence structures can be found in Sections 3.2.4 and 3.2.5.

One of the key findings from the literature described thus far is that the statistical methodology proposed for the evaluation of surrogate endpoints, and in particular the two-stage approach, must be re-defined for each different type of surrogate and true endpoint, and combination thereof. Methods that are suitable for continuously distributed endpoints are not immediately transferable to time-to-event or ordinal outcomes, and whilst measures of surrogacy continue to take values within $[0, 1]$ it is difficult to know whether such measures are comparable. As a result, further approaches proposed for the evaluation of surrogate endpoints have adopted a more unified framework, to enable their use with any types of surrogate and true outcomes. These developments are described in the forthcoming section.

2.4 Meta-Analytic Unified Measures

The development of unified measures for meta-analytic evaluation of surrogate endpoints is based on key concepts of generalised linear models and the likelihood ratio test statistic for comparing them, and so these are first defined below.

- **Generalised Linear Models** are a class of models that relate the mean (μ) of an outcome variable Y to a linear combination of covariates ($\eta = \beta^T X$) via a link function, g, so that $\mu = E(Y) = g^{-1}(\eta)$. The benefit of such a class of models is the flexibility in choice of link function, allowing use of the models for many different

outcome types.

- The **Likelihood Ratio Test** is a method that is used to compare the likelihood values of nested models, thereby determining whether the inclusion of additional covariates can significantly improve the fit of a model to observed data. Suppose that two models exist; a 'reduced' model (H_0) and an 'unrestricted' model (H_A), where the 'reduced' model is a nested version of the 'unrestricted' model and contains only a subset of the covariates of that full model. Let L_0 and L_A be the values of the likelihood for models H_0 and H_A respectively, calculated under the maximum likelihood estimates of the model parameters, or covariate coefficients, $\widehat{\beta}$. Then, the likelihood ratio test statistic, $G = -2\ln\left(\frac{L_0}{L_A}\right)$, can be compared to the chi-squared distribution to provide a test of whether there is a significant difference in the value of the likelihood through inclusion of the additional covariate(s). The degrees of freedom for this chi-squared distribution is equal to the difference in the number of covariates between the two models. A higher value of G corresponds to a higher likelihood from the alternative model (H_A), and therefore evidence to suggest that the additional covariates improve the model fit.

These concepts are highly relevant to the setting of surrogate endpoint evaluation. When considering individual-level surrogacy, it would be of interest to determine whether a model of the true outcome could be improved through inclusion of both treatment and the surrogate outcome as covariates, as compared to a model containing treatment only. At the trial-level, it would be of interest to determine whether a model of the treatment effect on the true endpoint could be improved through inclusion of the treatment effect on the surrogate outcome as a covariate. One key benefit of using the likelihood ratio test together with generalised linear models for this purpose is that any type of endpoint, be it continuous, binary or longitudinal, can be accommodated through the choice of model. These ideas were therefore used to develop a unified framework for surrogacy evaluation.

One of the first unified approaches to the assessment of surrogate endpoints, the Vari-

ance Reduction Factor, was proposed by Alonso et al. (2003) for the specific setting of longitudinal surrogate and true endpoints, both being measured multiple times over the course of the clinical trial. An extension to this measure that can handle different types of endpoints was subsequently proposed by Alonso et al. (2006), the Likelihood Reduction Factor (LRF).

Through the use of two generalised linear models, one containing a covariate of treatment only, and one containing both treatment and the surrogate, the LRF was first proposed as a measure of individual-level surrogacy based on the likelihood ratio test between the two models as

$$LRF = 1 - \frac{1}{N}\sum_i \exp\left(-\frac{G_i^2}{n_i}\right),$$

where G_i^2 denotes the likelihood ratio test statistic for trial i containing n_i patients, with a total of N trials available for analysis. The LRF defines a unified measure of individual-level surrogacy that is consistent across all endpoint types, lies within $[0, 1]$ and takes a value of zero when S and T are independent. However, O'Quigley and Flandre (2006) argue that the LRF, despite being applicable to all endpoints, does not adequately account for censoring when the surrogate and true endpoints are of a time-to-event structure. A measure based on the same quantities, but weighted according to the number of events per trial rather than the sample size per trial is therefore proposed (LRF-a), which can also be estimated using standard software:

$$LRF_a = 1 - \frac{1}{N}\sum_i \exp\left(-\frac{G_i^2}{k_i}\right),$$

where k_i represents the number of events observed in trial i. O'Quigley and Flandre (2006) recognise that the LRF and LRF-a are specific examples of a wider group of statistical concepts, the use of which could provide further measures which would be applicable for the assessment of surrogate endpoints. This framework is therefore described in the next section.

2.4.1 Information Theory

The underlying concept of LRF and LRF-a measures lies within a framework of statistical methodology focused on information gain, or explained randomness (O'Quigley and Flandre, 2006). Within this framework, the distance between two nested models can be measured in terms of the 'information gain' that comes from inclusion of covariates in one model but not the other. Such information gain provides quantification of how much uncertainty, or randomness, in an outcome can be explained by the addition of covariates into the model, and therefore a measure of how much uncertainty can be removed by accounting for a given covariate.

Such interpretation is highly relevant for the evaluation of surrogate endpoints; an endpoint that can capture a large amount of the uncertainty in outcome could be considered a good candidate as a surrogate endpoint. Further, the level of uncertainty in treatment effect on T that can be removed through knowledge of the treatment effect on S would provide information as to the reliability of a surrogate at the trial-level (Alonso and Molenberghs, 2007). The underlying concepts of the information theory approach are now described in terms of two random variables A and B, and subsequently described in the context of surrogate and true endpoints S and T.

The fundamental concept of information theory lies in the entropy of a random variable, which provides a measure of the uncertainty of that random variable (Shannon, 1948). For continuous variables B and A with density functions f_B and f_A respectively, the differential entropy of B and conditional entropy of $B|A$ are defined as

$$\begin{aligned} h(B) &= \int_b f_B(b) \log f_B(b) db \\ h(B|A) &= \int_b f_{B|A}(b|A=a) \log f_{B|A}(b|A=a) db, \end{aligned}$$

where the conditional entropy $h(B|A)$ provides a measure of the uncertainty in B that remains after accounting for another random variable A. Using these quantities, Shannon (1948) further defines the concept of entropy power (EP), a measure which allows

comparison of the entropy of multiple random variables;

$$EP(B) = \frac{1}{(2\pi e)^n} e^{2h(B)},$$

with n denoting the number of units in the sample. Based on these ideas, Alonso and Molenberghs (2007) define a general measure that describes how much of the uncertainty in a given outcome, B, can be removed through knowledge of another random variable, A;

$$R_h^2 = \frac{EP(B) - EP(B|A)}{EP(B)}.$$

An alternative representation of R_h^2 can be described using the difference between the entropy of B and the conditional entropy of $B|A$ as

$$R_h^2 = 1 - e^{-2I(A,B)},$$

where $I(A, B) = h(B) - h(B|A)$ denotes the mutual information between A and B; the amount of uncertainty in B that is removed when A is known. When A and B are independent, knowledge of A is irrelevant to the uncertainty in B and so $h(B|A) = h(B)$ and the mutual information is zero, leaving $R_h^2 = 0$. If A contains a large amount of information about B, then knowledge of A reduces much of the uncertainty in B, leading to $h(B|A)$ close to zero. In this case, $R_h^2 \approx 1$.

The interpretation of such a measure is then directly applicable to the evaluation of surrogate endpoints, where the aim is to understand the amount of uncertainty in T that can be explained by S at the individual-level, or, at the trial-level, the amount of uncertainty in treatment effect on T that can be explained by the treatment effect on S. In the remainder of this thesis, $R_{h,i}^2$ will be used to denote individual-level surrogacy, and $R_{h,t}^2$ trial-level surrogacy, based on the information theory approach. Alonso and Molenberghs (2007) note that $R_{h,i}^2$ can be estimated using LRF, however, given the criticism of this approach by O'Quigley and Flandre (2006) when using censored data, alternative approaches to estimation of $R_{h,i}^2$ for time-to-event outcomes were explored by Pryseley

et al. (2011). This investigation is highly relevant to the remainder of this thesis, and is therefore described in detail next.

2.4.2 Time-to-Event Endpoints

Following the initial proposal of a more general information-theory approach to surrogacy evaluation by Alonso and Molenberghs (2007), the extension to time-to-event endpoints was immediate. However, the criticism that the measure, in the proposed form, did not adequately account for the presence of censoring led to further work to determine which of a selection of methods could be considered most appropriate for censored data. In order to investigate this, Pryseley et al. (2011) conducted a simulation study comparing three different methods to estimate $R^2_{h,i}$ in the assessment of individual-level surrogacy. Of note is that the theory behind these measures is the same, the only difference is in how the mutual information, $I(S,T)$, is estimated in the surrogacy measure $R^2_{h,i}$.

The first measure that was considered was the LRF in the originally-proposed form, weighting by the trial size. The second measure was LRF-a, with weighting based on the number of events per trial rather than the number of patients (O'Quigley and Flandre, 2006). The third and final measure included in their simulation study was a measure proposed by Xu and O'Quigley (1999) based on a dependence measure for proportional hazards models, which can be denoted by R^2_{XOQ} in line with Pryseley et al. (2011). Results of the simulation study of Pryseley et al. (2011) demonstrated that LRF and LRF-a performed poorly across all scenarios investigated, and so these measures are not further discussed herein. R^2_{XOQ} was recommended for further use, and is therefore described in detail here, with further investigation of this measure presented in Chapter 4.

To describe R^2_{XOQ} for the assessment of trial-level surrogacy, it is first necessary to provide some notation and assumed models. First, a Cox proportional hazards model is used to define the hazard function for the outcome T as

$$\lambda(t|X(t)) = \lambda_0(t)\exp(\beta X(t)),$$

where $\lambda_0(t)$ denotes the baseline hazard at time t, β denotes a vector of covariate coefficients, which are considered constant over time, and $X(t)$ denotes a vector of covariate values for each patient. For these models, $X(t)$ includes treatment (Z_{ij}), the surrogate endpoint (S_{ij}), and any other relevant covariates. In future descriptions, only covariates of treatment and surrogate outcome will be discussed; the addition of other covariates does not impact the procedure for calculation of R^2_{XOQ}.

When aiming to assess the value of a potential surrogate endpoint at the individual-level, it is of interest to understand whether knowledge of that surrogate outcome provides sufficient, reliable information on the true outcome to warrant future use. In the context of information theory, comparison of a model of the true outcome containing covariates treatment only, versus a model containing treatment and the surrogate outcome as covariates, would provide an estimate of how much the uncertainty in the true outcome could be reduced by accounting for the surrogate outcome.

To estimate R^2_{XOQ}, two models are assumed. Splitting the coefficient vector $\beta = (\beta_1, \beta_2)$ to represent covariate coefficients for treatment and surrogate respectively, the 'null' model, H_0, corresponds to a model where β_2 is assumed to be zero, such that the surrogate has no effect on T. An 'alternative' model, H_1, corresponds to a model where β_2 has no restriction:

$$H_0 : \lambda(t|Z(t)) = \lambda_0(t)\exp(\beta_1 Z), \tag{2.2}$$

$$H_1 : \lambda(t|Z(t)) = \lambda_0(t)\exp(\beta_1 Z + \beta_2 S). \tag{2.3}$$

Using the notation of Pryseley et al. (2011), if the true value of $\beta = (\beta_1, \beta_2)$ were denoted by β_0, and $\widetilde{\beta_1}$ denotes the value maximising the likelihood of the model with respect to β under model H_0, then the measure R^2_{XOQ} between models H_0 and H_1 is defined as

$$R^2_{XOQ} = 1 - \exp(-\Gamma(H_1, H_0; \beta_0)),$$

where

$$\Gamma(H_1, H_0; \beta_0) = 2\int_T \int_Z \log\left[\frac{g(z|t;\beta_0)}{g(z|t;\widetilde{\beta_1})}\right] g(z|t;\beta_0)dzdF(t), \tag{2.4}$$

with $g(z|t,\beta)$ denoting the conditional distribution of $Z|T$ and $F(t)$ the marginal distribution function of T. In order to estimate R^2_{XOQ} in practice, it is necessary to estimate these conditional and marginal distributions.

To enable estimation over the entire domain of T, the values of the marginal and conditional distributions are required at each event time, t_k. For the marginal distribution of T, the distribution that ignores all covariates, the Kaplan-Meier estimate of survival is proposed (Kaplan and Meier, 1958), where the marginal distribution $F(t)$ is then defined as the step, or jump, in the Kaplan-Meier function at each event time, $W(t_k)$.

For the conditional distribution of $Z|T$ at each event time, Xu and O'Quigley (1999) propose to use the conditional probability of patient j having an event at time t_k given their covariate values at that time, which in a single trial (i.e. no subscript i for trial) is defined as

$$\pi_j(t_k,\beta) = \frac{Y_j(t_k)\exp(\beta Z_j)}{\sum_{l=1}^{n} Y_l(t_k)\exp(\beta Z_l)}, \tag{2.5}$$

where the sum over $l = 1, ..., n$ denotes all patients at-risk of an event at time t_k, and $Y_j(t_k)$ denotes an indicator variable to determine whether patient j is at risk of an event at time t_k. The product of these values over all event times forms the partial likelihood of the Cox proportional hazards models which are used to estimate β_1 and β_2 in equations (2.2) and (2.3) (Cox, 1972). Finally, $\Gamma(H_1, H_0; \beta_0)$ is then estimated from a single trial, i, using:

$$\widehat{\Gamma}(H_1, H_0; \widehat{\beta}_0) = 2\sum_{k=1}^{K} W(t_k) \sum_{j=1}^{n} \pi_{ij}(t_k; \widehat{\beta}_0) \log\left[\frac{\pi_{ij}(t_k; \widehat{\beta}_0)}{\pi_{ij}(t_k; \widetilde{\beta}_1)}\right],$$

where $k = 1, ..., K$ denote the number of events for the true outcome, $W(t_k)$ the jumps in the Kaplan-Meier function at time t_k, and $\pi_{ij}(.)$ the probability of individual j in trial i being the patient having the event at t_k.

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 easily by re-calculating the measure using the confidence limits for the surrogate covariate coefficient (β_L, β_U), i.e. $[R^2_{XOQ}(\beta_L), R^2_{XOQ}(\beta_U)]$. In a meta-analytic setting, R^2_{XOQ} can be calculated

for each trial and a weighted estimate provided for an overall measure. Xu and O'Quigley (1999) note that the measure is dependent on the total duration of follow-up of the trial, and propose a corrected version that divides by the sum of $W(t_k)$ values over all events. Further, when covariates are categorical and have very few levels, R^2_{XOQ} is bounded by a number less than one, although this is considered to be irrelevant except for very high values of R^2_{XOQ} (Pryseley et al., 2011).

The description of the information theoretic approach has thus far been limited to individual-level surrogacy, however the approach also has a useful interpretation at the trial-level. Such a measure would estimate how much of the uncertainty in treatment effects on the true endpoint can be reduced through knowledge of treatment effects on the surrogate endpoint. Alonso and Molenberghs (2007) note that an information theoretic measure of trial-level surrogacy reduces to the R^2_{trial} measure of Buyse et al. (2000) when a linear relationship between treatment effects is assumed, and therefore immediately has an interpretation from the information theory perspective. This finding leads to the focus of the work by Pryseley et al. (2011) being restricted to individual-level surrogacy.

2.4.3 Other Relevant Areas of Investigation

Motivation for the development of the information theory approach to evaluating surrogacy stemmed from the challenges faced when attempting to use previously proposed measures of surrogacy, such as the two-stage meta-analytic copula method. The need for endpoint-specific modelling structures, and the complexity of the numerical processes involved in some of those approaches, are not considered to be an issue for the information theory method. However, there remain some questions that require further examination.

Pryseley et al. (2011) conducted a detailed simulation study to examine the performance of R^2_{XOQ} as an estimate of $R^2_{h,i}$ for time-to-event data, which highlighted a number of aspects worthy of further discussion. Firstly, Pryseley et al. (2011) did not assess the performance of R^2_{XOQ} when based on a meta-analysis of multiple datasets. The measure

was assessed using simulations for just one clinical trial, and so the performance of the measure when based on multiple studies has not been investigated.

Additionally, as for the two-stage meta-analytic copula method, the simulation study of Pryseley et al. (2011) considers a surrogate that is censored by the true endpoint (T-dependent censoring of S), and does not explore the impact of the use of surrogates that contain data directly related to the true endpoint, such as progression-free survival. Interestingly, it is noted that when such an endpoint is being assessed, the information theory measure can be directly applied to modelling of the post-progression survival time, $T - S$, rather than the overall survival time. Such an approach would enable the dependency between endpoints to be accounted for without the need for a change in modelling approach. This topic remains unexplored in the literature.

In addition to the simulation study presented by Pryseley et al. (2011), further investigation was conducted by Pryseley (2009), where the sensitivity of R^2_{XOQ} to the proportional hazards assumption was explored. Since the proposed information theory method utilises proportional hazards models, it is of interest to understand the impact of use of the approach when this assumption is violated. Results of this study showed that, encouragingly, R^2_{XOQ} performs well when there is low to moderate censoring and large sample sizes, therefore supporting the use of the information theory method in practice.

2.4.4 General Limitations of Surrogacy Evaluation Measures

In addition to limitations and outstanding questions specific to the two surrogacy evaluation methods described in Sections 2.3.4 and 2.4.3, there are a number of more general issues that are applicable to multiple surrogacy evaluation approaches.

Both the two-stage method for assessing surrogacy and the information theory approach, regardless of whether the outcome is continuous, binary or time-to-event, provide individual and trial-level measures of surrogacy that lie within $[0, 1]$, with 'poor' surrogacy demonstrated by values close to zero, and 'good' surrogacy by values close to one.

However, there is currently no established threshold that has been considered to truly demonstrate that a surrogate is reliable for use. There is therefore the potential for substantial subjectivity as to the level of evidence that must be demonstrated for a surrogate to be established. It is also difficult to know whether surrogacy measures from different endpoint types are comparable and have the same interpretation of the strength of relationship.

In a recent meta-analysis evaluating a potential surrogate for a type of blood cancer, Shi et al. (2017) estimated R^2_{trial} using two methods; weighted least squares regression and the two-stage copula approach. A pre-specified threshold for 'success' was defined as one of these estimates of R^2_{trial} being ≥ 0.80, with a lower bound of the 95% confidence interval > 0.60, and both values of $R^2_{trial} \geq 0.70$. This appears to be the first example in the literature where pre-specification has been made, and may set the precedent for future requirements for surrogates to be accepted by regulatory authorities. Nevertheless, there are likely to be arguments that such thresholds are extremely high, particularly when surrogacy evaluation is based on a small number of trials where it can be difficult to achieve such accurate predictions.

A further limitation, or challenge, is the need for large amounts of patient-level data to be available to ensure that conclusions are representative for the widest possible range of future clinical settings. In many cases, such data are very difficult to obtain, in particular for less severe diseases where patients can have a good prognosis and the time to observation of clinical events of interest (e.g. disease progression or death) are very long. This has led to many applications of surrogacy evaluations to small sample settings where only a small number of clinical trials are available. In such settings, a common approach is to split the studies into smaller subgroups, a topic that is discussed further in the next section.

Finally, a difficult hurdle to overcome is gaining consensus amongst clinical and statistical researchers, as well as regulatory authorities, on the acceptability of the proposed statistical methodology for use in practice. Whilst the two-stage meta-analytic copula

method has commonly been used in surrogacy evaluations, as in the examples in Section 2.3.3, other methods have been described as a preferred approach, but not often used in practice (Ensor et al., 2016). Currently, there is no general agreement on which methodology should be applied in the assessment of surrogacy. Further research and collaboration between research groups is therefore warranted.

2.5 Analysis of Trials versus Centres

Ideally, surrogacy evaluation would be conducted using multiple, large, clinical trial datasets, so that estimates of model parameters are reliable and subsequent surrogacy measures are robust. However, it is often the case that such large databases are not available, and in these circumstances some researchers have opted to split the available studies into subgroups (e.g. Dimier et al. (2015) who split studies based on geographical location of the investigational sites).

Such splitting of clinical trial data has been a commonly used approach to increase the number of data points available for analysis, and has therefore been the subject of investigation by a number of researchers. Shi et al. (2011) suggest that splitting studies into arbitrary subgroups, or by investigational centre, leads to the need for consideration of the balance between gain in precision and loss of accuracy, the effects of which are not well understood. For the case of continuous surrogate and true outcomes, Burzykowski et al. (2005) explore a number of strategies to assess the performance of modelling at trial and centre levels, when the level of association at these levels is similar, or varies. When similar levels of association are assumed, the measures under investigation were found to provide reasonable results, whereas when the association levels differ, the use of centres in analysis was found to lead to biased results. However, in some circumstances, notably when the variability in centre-level treatment effects was assumed to be lower than that of trial-level treatment effects, the estimated levels of association based on centres appeared similar to the true trial-level association, thereby providing some rationale for the use of

centre-level analysis in practice.

Further to this, Renfro et al. (2014) conducted a detailed examination of the performance of the two-stage meta-analytic copula method and weighted least squares regression when based on subgroups of trials, for time-to-event endpoints. Similar to Burzykowski et al. (2005), estimation using centres was determined to be acceptable when the variability between treatment effects at a centre-level is similar to or lower than that of treatment effects at a trial-level. Since these scenarios are considered unlikely in practice, a simulation study was conducted to assess the bias in estimation of R^2_{trial} through use of centres as the units for analysis. The authors consider a large range of scenarios, at a high-level and at a more focused level. The high-level investigation considers some variation of the true trial, centre and patient-level association, censoring, and numbers of trials, centres and patients. However, the majority of these scenarios considered high sample sizes, varying from a total of 1500 to 60,000 patients, and are therefore less relevant to the subject of limited data. More focused scenarios considered smaller sample sizes of 1-5 trials, each with 5-20 centres containing 10-50 patients, which is more relevant to the interest of this thesis, but these scenarios did not consider any variation in association; all three levels (trial, centre, patient) were held fixed at 0.9, with no censoring.

The results of this simulation study indicate that whilst centre-level analysis could be considered reliable under some circumstances, it quite often leads to bias. Recommendations as to the most appropriate unit for analysis are therefore made based on the number of trials available; when there exist at least ten trials, the analysis should be conducted at a trial-level only, whereas when there are 5-9 trials available, the centre-level analysis can also be used as supportive evidence of surrogacy. When a surrogacy evaluation is to be based on data from only 3-4 trials, Renfro et al. (2014) recommend to conduct analysis both at a trial-level, centre-level across all trials, and centre-level within each trial individually, with varying combinations of the results of these analyses used to determine the final conclusion. Overall, it is considered that surrogacy evaluation based on a limited number of trials is possible, with careful consideration of the available data.

Despite the large number of scenarios investigated by Renfro et al. (2014), the majority were based on the existence of large meta-analytic databases, which offers little insight into the perspective of an individual pharmaceutical company which has very little data on which to base a surrogacy assessment. The more focused scenarios provide some level of insight, but have only been considered for truly high-levels of association, with no consideration of whether centre-level analysis could be reliable when the association levels are reduced. From the analysis of larger sample sizes, it was found that when the true trial-level surrogacy is low (0.2) or moderate (0.5) and centre-level association is high, estimation of R^2_{trial} using centres could be biased upwards. Further investigation of the impact of smaller sample sizes on this finding is therefore needed.

2.6 Other Surrogacy Approaches

The two methods for the evaluation of surrogacy described in detail in previous sections are those that are used most commonly in practice (two-stage meta-analytic copula method) or recommended as a preferred approach (information theory (Ensor et al., 2016)). However, there are a number of alternative measures that have been proposed in the literature. In this section, alternative approaches are briefly described for completeness. First, other measures proposed specifically for time-to-event endpoints are briefly introduced. Subsequently, the surrogate threshold effect (STE), which provides a convenient measure of surrogacy that is easy for clinicians to interpret, is described. Finally, a framework constructed to allow a causal interpretation of potential surrogate endpoints is briefly discussed, which is a relatively new field in the area of surrogate endpoint research and has yet to be established as a standard approach.

2.6.1 Time-to-Event Endpoints

Recognising that the assumptions of the two-stage meta-analytic copula method, namely symmetry of endpoints, may not always hold, Ghosh et al. (2012) propose an alternative

approach that accounts for endpoint ordering by modelling the region where the surrogate endpoint occurs before the true endpoint ($S \leq T$). The key difference between this approach and the previously described surrogacy approaches is in the handling of the time-to-event surrogate endpoint. For endpoints such as progression-free survival that capture multiple event types, the method of Ghosh et al. (2012) separates these events, and considers that the surrogate is dependently censored by the true endpoint. The construction of composite endpoints, where multiple events are captured, is then removed.

Ghosh et al. (2012) consider the main advantage of the semi-competing risks approach to be that it is possible to study the time to each individual event type separately, without the need for creation of composite endpoints. Whilst this approach does avoid the need for the symmetry assumptions, it is of concern that the surrogate endpoint being evaluated differs from the composite endpoint that is well-understood and is desired for use in subsequent clinical trials. Separation of the endpoint into the two components (disease progression and death) is not considered an appropriate approach for settings where the combined endpoint is recognised by clinicians and regulators as a clinically relevant measure. Hence, this approach is not considered further in this thesis.

Parast et al. (2017) consider the proportion of treatment effect explained when calculated from censored data, specifically for the setting where the true clinical endpoint may be observed earlier than the surrogate endpoint, thereby introducing missing data into the surrogate endpoint evaluation. This scenario could be considered to further examine the impact of surrogacy evaluation when the surrogate is a composite of an intermediate disease state and the true clinical outcome. However, this approach is founded in an area of surrogacy evaluation literature that is relatively new and remains largely unestablished in practical evaluations of surrogacy (see Section 2.6.3), and is therefore not considered further here.

2.6.2 Surrogate Threshold Effect

The majority of measures proposed for the evaluation of surrogate endpoints result in a value lying within $[0, 1]$, with values close to zero being evidence of a poor surrogate, and values close to one demonstrating a good surrogate. Concerned that clinicians and regulatory agencies may face difficulty in interpreting such a value in the context of the potential patient benefit, Burzykowski and Buyse (2006) propose an additional measure that has a clear clinical interpretation, the Surrogate Threshold Effect (STE).

To construct the measure, Burzykowski and Buyse (2006) use an estimated value of R^2_{trial} from a (meta-analytic) surrogate endpoint evaluation, together with an observed treatment effect on S in a new clinical trial, to derive a prediction interval for the predicted treatment effect on T in that new clinical trial. The width of this prediction interval defines how large the treatment effect on S would need to be to ensure a statistically significant difference between treatments on the true endpoint. The treatment effect on S that would provide a lower bound of this prediction interval to be greater than some 'null' value, indicating no treatment effect, then provides the minimum treatment effect that needs to be observed on the surrogate to have confidence that the predicted treatment effect on T would be statistically significant.

The STE provides a clinically interpretable measure of the treatment effect required from a surrogate endpoint to be confident that a significant treatment effect on the true endpoint would also exist. If this required treatment effect on S were too small to be considered clinically meaningful, or indeed if it were felt to be so strong that it was infeasible to achieve, then this provides an additional level of information that cannot be captured by the previously described individual and trial-level surrogacy measures. However, since the STE can only be estimated when the nature and strength of relationship between S and T is reliably established, it is not considered to be a replacement for the previously described surrogacy measures, but rather provides additional information to aid interpretation. Hence, it is not discussed further in this thesis.

2.6.3 Causal Inference

Some authors have argued that many of the proposed statistical methods for evaluation of surrogate endpoints are flawed, in that they cannot be used to establish a causal link between treatment and long-term clinical outcome (Frangakis and Rubin, 2002). By adjusting for the surrogate outcome, which is measured at some point after treatment has started, it is not possible to conclude that differences in the true outcome are a result of treatment alone. As a result, a further branch of statistical methodology aimed at surrogacy evaluation has developed more recently, based on alternative approaches that maintain a causal interpretation.

Two general frameworks for the assessment of surrogates using causal association have been proposed by Frangakis and Rubin (2002) and Robins and Greenland (1992), with an investigation into the relationship between these proposals conducted by VanderWeele (2008). Further comparisons of surrogacy evaluation frameworks allowing causal interpretation were conducted by Joffe and Greene (2009).

More recently, Alonso et al. (2015) investigate the relationship between the causal effect and meta-analytic frameworks for surrogacy evaluation, comparing and discussing the underlying theoretical components of both paradigms, and applying measures from each to a case study. Overall, it is concluded that the meta-analytic approach has a number of advantages as a method for evaluating surrogate endpoints. Firstly, whilst not having a causal interpretation for the assessment of individual-level surrogacy, the assessment of trial-level surrogacy maintains such an interpretation as it does not adjust for post-treatment variables and is therefore not subject to this bias. Further, it is potentially more useful in practice due to greater appeal and simplicity, particularly for regulatory authorities.

Whilst a number of causal approaches have been proposed for both single-trial and meta-analytic surrogate evaluation, Ensor et al. (2016) note that such measures remain "in their infancy", with no agreement on which may be the most suitable or appropriate

for future use. The proposed approaches suffer from the need for strict and often unverifiable assumptions in order to estimate parameters, with poor estimation under even minor violation of these assumptions. As such, there is currently no consensus as to how best to estimate surrogacy using these methods, and no standard implementation. As a result, none of these measures are considered further in the context of the research presented in this thesis. Subsequent chapters of the thesis therefore focus on the two previously described measures, the two-stage meta-analytic copula method, and the information theory method.

Chapter 3

Two-Stage Meta-Analytic Copula Method for Evaluating Time-to-Event Surrogate and True Endpoints

3.1 Introduction

In Section 2.3.2, the two-stage meta-analytic copula method for assessing surrogacy of time-to-event surrogate and true endpoints was described, alongside identification of the limitations and some outstanding questions related to this approach. In this chapter, an attempt to address these questions is made through the use of a simulation study. Data are simulated according to two different underlying structures, to investigate the sensitivity of the approach to correctly and incorrectly specified models. A wide range of scenarios are considered, including consideration of different surrogate endpoints, different strengths of individual-level and trial-level surrogacy (low, medium and high, for each) and different sample sizes (both varied numbers of patients within each trial, and varied numbers of trials). The impact of censoring is also considered through the proportion of censored

observations, and the range of observed treatment effects on the true endpoint is also varied. This work has also been published in Pharmaceutical Statistics (Dimier and Todd, 2017).

Whilst the performance of the two-stage meta-analytic copula method has been studied via simulation by Burzykowski (2001) (described briefly also by Burzykowski et al. (2001)), a limitation of that study is that there was no exploration of one of the most commonly used surrogate endpoints, progression-free survival. The use of PFS violates the symmetry assumption of the copula modelling approach, and the impact of this on parameter estimation remains unknown. Therefore, the impact of the use of PFS as the potential surrogate is investigated within this chapter.

A second limitation with interpretation of the study conducted by Burzykowski (2001) is that the simulations were constructed using reasonably large sample sizes (10-20 trials, each consisting of 50-200 patients). In the setting of interest in this thesis, the assessment of surrogacy is assumed to be undertaken by individual pharmaceutical companies who have very limited data from individual clinical development plans available. One example of this can be seen in the work of Dimier et al. (2015), who attempt to assess surrogacy using data from only three trials. The simulation study described here therefore assesses the performance of the two-stage meta-analytic copula method when there exist very limited data.

Finally, whilst the aim is to identify surrogate endpoints that have high levels of predictive strength, it is also important that any statistical methodology can correctly identify those with truly low levels of predictive strength. This is increasingly important when small sample sizes are used, since it could be hypothesised that this would lead to lower precision in parameter estimation, potentially leading to less reliable conclusions. Burzykowski (2001) investigated only medium (0.5) to high (0.9) levels of association, and it remains unclear whether the two-stage meta-analytic copula method can reliably identify poor surrogates. Lower levels of both trial-level and individual-level association are therefore explored in the simulation study described below.

3.2 Simulation Study

3.2.1 Choice of Data Generation Procedure

One of the most important factors in setting up a simulation study is ensuring that the individual and trial-level association can be accurately controlled, such that the input values can be used as a reference against which the estimates provided by the modelling can be compared. In order to achieve this, Burzykowski (2001) generated data using a (Clayton) copula function, which allows for controlling of individual-level association through the copula dependence parameter. Whilst this allows for consistency between the underlying structure of the simulated data and the assumptions of the two-stage meta-analytic copula method (also used with the Clayton copula), there could be concern that this may lead to overly precise estimation of model parameters, as compared to real-life application of the methodology where data may deviate from the assumed dependence structure.

To address these concerns, two different copula functions, with two different underlying data structures, are considered for the simulation study described herein. First, the Clayton copula is used to generate data under 'ideal' circumstances where the dependence structure matches perfectly that assumed by the surrogacy evaluation approach. Second, to assess the impact of violating the assumed dependence structure, data generation is also conducted using a Gumbel copula, which assumes a very different dependence structure. In both cases, the two-stage meta-analytic copula method employs a Clayton copula modelling approach, so that an assessment of the impact of using a correctly versus incorrectly specified model can be made. Algorithms used to generate data are detailed in Sections 3.2.4 and 3.2.5 for the Clayton and Gumbel functions, respectively.

3.2.2 Selection of Surrogate Endpoints

With respect to potential surrogate endpoints, two different time-to-event endpoints that are commonly used in late-stage (Phase III) oncology clinical trials are considered as replacement endpoints for a true endpoint of overall survival. In a clinical trial setting, these endpoints are:

- **Time-to-Progression (TTP)**, defined as the time from entry into a clinical trial (e.g. from date of randomisation) until observation of disease progression. Patients who do not experience disease progression are censored at the time that their disease was last assessed by the treating physician.

- **Progression-Free Survival (PFS)**, defined as the time from entry into a clinical trial until the patient experiences disease progression or death, whichever occurs first. Patients who do not experience disease progression or death during the period of observation are censored at the time that their disease was last assessed by the treating physician.

- **Overall Survival (OS)**, defined as the time from entry into a clinical trial until death. Patients who remain alive at the end of follow-up are censored at the time they were last known to be alive.

The two endpoints being assessed as potential surrogates, TTP and PFS, are therefore very similar, with the exception that PFS also includes death as an event of interest. Since the true endpoint of interest here is OS, both TTP and PFS must be truncated at the time of death, since no further follow-up is possible. However, the key difference between the two surrogates is that an event of death censors TTP, whereas PFS includes this as an event of interest. Therefore, for TTP it is assumed that the time to disease progression *could* be longer than the time to death, but death precludes observation of the progression. In contrast, PFS can never be longer than OS, since death is included as an event. This is a critical observation, as the endpoint of PFS then violates the copula model assumption of

symmetry of endpoints. The simulation study described here therefore includes both TTP and PFS as surrogate endpoints and allows an assessment of the impact of this violation, something which was not considered by Burzykowski (2001).

3.2.3 Defining Simulation Parameters

As described above, the data generation used in this simulation study was based on two surrogate endpoints, TTP and PFS, and two copula functions, Clayton and Gumbel. Further to this, a number of other variables were considered. Firstly, the number of trials and respective sample sizes were selected to represent the setting of individual pharmaceutical companies, who have limited data from their own clinical development plans only. Varied censoring proportions were assumed, to assess the sensitivity of the two-stage meta-analytic copula method to the amount of censoring in the data. Finally, previous studies have determined that the surrogacy evaluation approach may perform better when the range of results across the clinical trials included in the analysis is wide (Burzykowski et al., 2005), meaning that there is more variability in baseline hazards and treatment effects on T. Whilst this is certainly important for meta-analyses including all available trials, regardless of outcome, it may not be so relevant for the setting of interest here, where pharmaceutical companies are likely to have consistent results across multiple trials to warrant further development of the molecules(s). Nonetheless, different ranges of treatment effects on OS were considered to assess the impact on the performance of the surrogacy approach, and this is described further in Section 3.2.4. A summary of all simulation parameters and selected values are shown in Table 3.1.

In order to assess the performance of the two-stage meta-analytic copula method across this range of scenarios, both individual-level surrogacy (denoted τ due to use of this parameter to represent R^2_{indiv}) and trial-level surrogacy (R^2_{trial}) were estimated for a total of 5,000 repetitions per scenario. Given the total number of 1296 different scenarios, and with each run taking approximately one second to complete, this was the largest

Table 3.1: Simulation Scenarios

Factor	Scenarios under simulation
Surrogate Endpoint	TTP, PFS
Data Generation	Clayton, Gumbel
Number of trials	4, 6
Number of patients per trial	80, 120, Mixed (50% each at $n = (80, 120)$)
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%
Range of treatment effects*, σ	0.1, 0.2

*Hazard ratios ranging $42\% - 203\%$ and $31\% - 238\%$ from the mean for $\sigma = 0.1, 0.2$ respectively.

number of runs that was considered to be feasible, taking approximately 1800 hours. Selecting the same number of runs across all scenarios ensured that each was examined with consistent precision, and since previous studies have examined the performance of the method under only 500 simulation runs, it is considered that 5,000 runs provides superior reliability. In order to apply the two-stage meta-analytic copula method to the generated datasets, code was taken from the website: `http://ibiostat.be/online-resources/online-resources/surrogate` and adapted to add the simulation steps described below. Testing of this code was conducted to ensure that results of previous simulation studies could be replicated, as well as testing to ensure that results of previous case studies reported in Burzykowski et al. (2005) could be replicated.

3.2.4 Clayton Copula Data Generation

The first copula model used to generate data according to the requirements listed in Table 3.1 is the Clayton copula, which is also used in the application of the two-stage meta-

analytic copula method within this thesis. An illustration of this copula for a value of $\tau = 0.5$ can be found in Figure 3.1, with a surface plot on the left and scatterplot on the right. The peak of the surface plot around $(0, 0)$, and the increased correlation in the scatterplot at this point, demonstrate that the model exhibits strong late dependence between endpoints, i.e. stronger dependence when values of both marginal survival functions are close to zero, reflecting longer values of S and T.

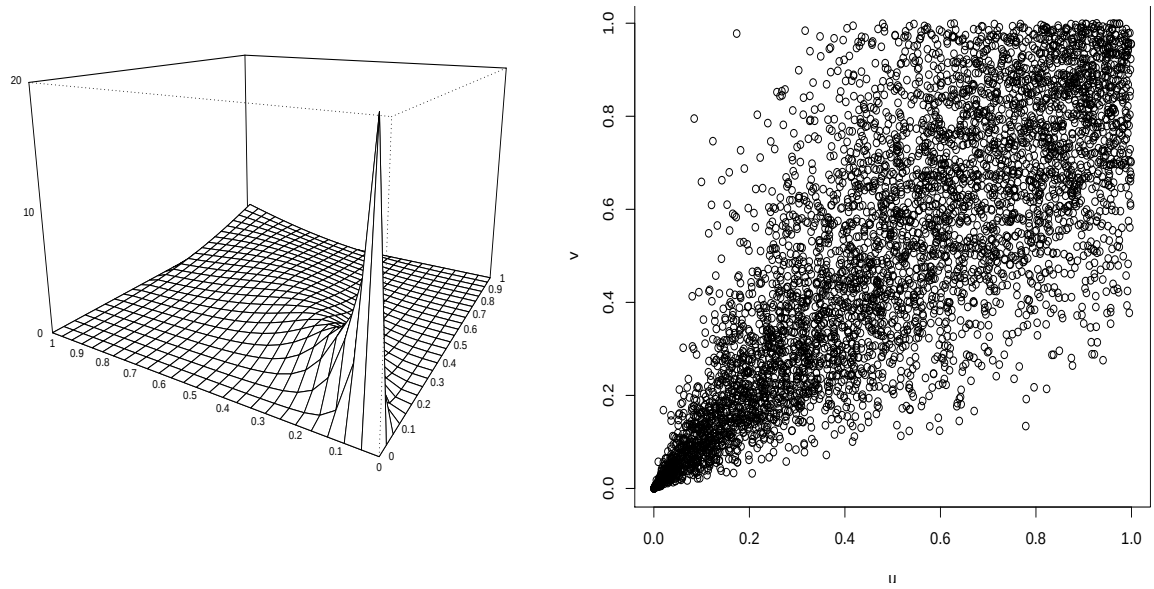

Figure 3.1: Clayton Copula Model with $\tau = 0.5$

The surrogacy evaluation approach, including the form of the copula function, is described in detail in Section 2.3.2, and briefly highlighted within this section where such explanation is helpful in detailing the structure and assumptions of the simulated datasets. As a reminder, the general form of the Clayton copula is defined as

$$C_{\theta_c}(u, v) = \left(u^{1-\theta_c} + v^{1-\theta_c} - 1\right)^{\frac{1}{1-\theta_c}},$$

where $C_{\theta_c}(u, v)$ represents the copula function with dependence parameter θ_c. This function leads to a joint survival function, $S(s, t)$ between surrogate (S) and true (T) endpoints,

$$S(s, t) = P(S_{ij} \geq s, T_{ij} \geq t) = C_{\theta_c}\{S_{S_{ij}}(s), S_{T_{ij}}(t)\}, \qquad s, t \geq 0,$$

where $S_{S_{ij}}(s)$ and $S_{T_{ij}}(t)$ denote the marginal survival functions for S and T, respectively, for patient j in trial i. In order to generate data according to the Clayton copula, three steps are taken:

1. Independent and identically distributed random variables U_{ij} and V_{ij} are generated from a Uniform$(0, 1)$ distribution for each patient j in trial i.

2. U_{ij} and V_{ij} are transformed according to the Clayton copula, with specified dependence parameter θ_c, to provide two uniformly-distributed variables (S^0_{ij} and T^0_{ij}) that are associated according to the dependence structure and strength specified.

3. S^0_{ij} and T^0_{ij} are transformed according to the selected marginal survivor distributions, to obtain two time-to-event outcomes with the required association. Further details of these three steps are provided below.

Once U_{ij} and V_{ij} are independently drawn from a Uniform$(0, 1)$ distribution, they are transformed to variables that have the specified dependence structure of the Clayton copula, using the conditional distribution method (Nelsen, 1999) (Step 2 above). This algorithm takes the first partial derivative of C_{θ_c}, with respect to one of the Uniform variables (U_{ij}), and inverts with respect to the remaining Uniform variable, (V_{ij}), to derive a transformation that provides the required copula dependence structure. For the Clayton copula, the first partial derivative is defined as

$$\frac{\partial C_{\theta_c}(U_{ij}, V_{ij})}{\partial(U_{ij})} = U_{ij}^{-\theta_c}\left(U_{ij}^{1-\theta_c} + V_{ij}^{1-\theta_c} - 1\right)^{\frac{\theta_c}{1-\theta_c}},$$

and equating this to V_{ij} leads to the following transformation:

$$\begin{aligned} V_{ij} &= U_{ij}^{-\theta_c}\left(U_{ij}^{1-\theta_c} + V_{ij}^{1-\theta_c} - 1\right)^{\frac{\theta_c}{1-\theta_c}} \\ \implies \left(U_{ij}^{\theta_c} V_{ij}\right)^{\frac{1-\theta_c}{\theta_c}} &= U_{ij}^{1-\theta_c} + V_{ij}^{1-\theta_c} - 1 \\ \implies V_{ij} &= \left(U_{ij}^{1-\theta_c} V_{ij}^{\theta_c^{-1}-1} - U_{ij}^{1-\theta_c} + 1\right)^{\frac{1}{1-\theta_c}}. \end{aligned} \tag{3.1}$$

Applying this transformation leads to two uniformly-distributed variables with joint distribution defined by the Clayton model, with strength of association denoted by θ_c:

$$\begin{aligned} S_{ij}^0 &= U_{ij}, \text{ and} \\ T_{ij}^0 &= \left(U_{ij}^{1-\theta_c} V_{ij}^{\theta_c^{-1}-1} - U_{ij}^{1-\theta_c} + 1\right)^{\frac{1}{1-\theta_c}}. \end{aligned}$$

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. As described in Section 2.3.2, Kendall's τ is instead used to measure the individual association between endpoints, and so this parameter is also used to control the individual association in data generation. For the Clayton copula, θ_c can be calculated directly from Kendall's τ using $\theta_c = \frac{1+\tau}{1-\tau}$, and so values of θ_c of 1.5, 3 and 9 achieve 'true' individual-level association of 0.2, 0.5 and 0.8 respectively.

Finally, the Uniform variables S_{ij}^0 and T_{ij}^0 must be transformed to be time-to-event variables (Step 3 above), S_{ij} and T_{ij}, according to the choice of marginal survivor functions. These marginal survivor functions are denoted as $S_{S_{ij}}(s)$ for the surrogate endpoint and $S_{T_{ij}}(t)$ for the true endpoint, where $S_{S_{ij}}(s) = P(S_{ij} \geq s)$ and $S_{T_{ij}}(t) = P(T_{ij} \geq t)$. To be consistent with Burzykowski et al. (2001), these marginal survivor functions are assumed to follow an exponential survival distribution, such that the final random variables S_{ij} and T_{ij} follow

$$\begin{aligned} S_{S_{ij}}(s_{ij}) &= \exp\{-s_{ij}\lambda_S \exp\{\mu_{Si} + (\alpha + a_i)Z_{ij}\}\}, \\ S_{T_{ij}}(t_{ij}) &= \exp\{-t_{ij}\lambda_T \exp\{\mu_{Ti} + (\beta + b_i)Z_{ij}\}\}, \end{aligned}$$

where λ_S, λ_T are baseline hazard functions specific to each endpoint (assumed to be constant), reflecting the expected survival for a patient with covariate values of zero, and α and β represent the treatment effects on the surrogate and true endpoints respectively, corresponding to the natural logarithm of the required hazard ratios in Table 3.1. The selection of values for λ_S, λ_T, α and β are discussed in Section 3.2.6. The assigned treatment group is represented by a binary covariate, Z_{ij}, taking values of zero (control

arm) versus one (experimental arm) and is drawn from a Bernoulli distribution assuming equal randomisation (parameter value 0.5). The remaining parameters, $(\mu_{Si}, \mu_{Ti}, a_i, b_i)$, represent trial-specific random effects, to account for variability across different clinical trials (further described below).

In order to convert the Uniform random variables, S^0_{ij} and T^0_{ij}, to be exponentially distributed, the following transformation is made:

$$S_{ij} = -\lambda_S^{-1} \exp\{\mu_{Si} + (\alpha + a_i) Z_{ij}\} \log\left(S^0_{ij}\right), \tag{3.2}$$

$$T_{ij} = -\lambda_T^{-1} \exp\{\mu_{Ti} + (\beta + b_i) Z_{ij}\} \log\left(T^0_{ij}\right). \tag{3.3}$$

The trial-specific random effects and treatment effects control the level of trial-level association between the surrogate and true endpoints, and are used to estimate R^2_{trial}. To control the strength of this association, the parameter vector $(\mu_{Si}, \mu_{Ti}, a_i, b_i)$ is assumed to follow a zero mean normal distribution with covariance matrix

$$D = \sigma \begin{pmatrix} 1 & \rho & 0 & 0 \\ \rho & 1 & 0 & 0 \\ 0 & 0 & 1 & \rho \\ 0 & 0 & \rho & 1 \end{pmatrix},$$

where σ is the parameter value from Table 3.1 chosen to control the level of variation in the random effects (and therefore the range of treatment effects in the trials), with larger values of σ leading to larger trial-specific random effects and therefore a larger range of treatment effects across simulated trials. The parameter ρ is chosen to be the square root of the required 'true' trial-level association (R^2_{trial}). The random effects are derived by generating vectors of independent standard-normal variables $(R_{1,i}, R_{2,i}, R_{3,i}, R_{4,i})$ and transforming using the Cholesky decomposition matrix

$$\begin{pmatrix} \mu_{Si} \\ \mu_{Ti} \\ a_i \\ b_i \end{pmatrix} = \sqrt{\sigma} \begin{pmatrix} 1 & 0 & 0 & 0 \\ \rho & \sqrt{1-\rho^2} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & \rho & \sqrt{1-\rho^2} \end{pmatrix} \begin{pmatrix} R_{1,i} \\ R_{2,i} \\ R_{3,i} \\ R_{4,i} \end{pmatrix}.$$

By following this 3-step algorithm, resulting datasets consist of two separate time-to-event endpoints (S and T) that have a joint distribution fitting the Clayton copula model, with strength of association defined by the copula parameter. Censoring is applied by drawing a random exponential variable C_{ij} and comparing to the values of S_{ij} and T_{ij}. Since the true endpoint is assumed to be 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 censoring is not applied. This is discussed further in Section 3.2.6.

3.2.5 Gumbel Copula Data Generation

As noted previously, since the Clayton copula function is chosen for use in the two-stage meta-analytic surrogacy approach, generating data according to this same function may be considered to be an 'ideal' case, where the underlying data structure fits perfectly the assumptions of the modelling approach. To assess the impact of this on the precision of parameter estimation, data were also generated using an alternative copula function, to enforce a different dependence structure than that assumed by the model.

The alternative copula function selected was the Gumbel copula, since this has a very different dependence structure to that of the Clayton copula and would therefore be a good candidate for assessing the impact of model misspecification. The dependency structure is different in that the Gumbel copula exhibits strong dependence for early event times, whereas the Clayton copula exhibits strong dependence for late event times. This can be seen in Figure 3.2, which demonstrates two peaks of the Gumbel copula function, with the greatest peak occurring where the marginal survival distributions are close to $(1, 1)$, and the scatterplot showing stronger correlation at this point. These reflect low event times, where the probability of survival remains high.

The general form of the Gumbel model for two random variables u and v is

$$C_{\theta_g}(u, v) = \exp\left[-\left\{(-\ln u)^{\frac{1}{\theta_g}} + (-\ln v)^{\frac{1}{\theta_g}}\right\}^{\theta_g}\right], \text{ for } 0 < \theta_g < 1.$$

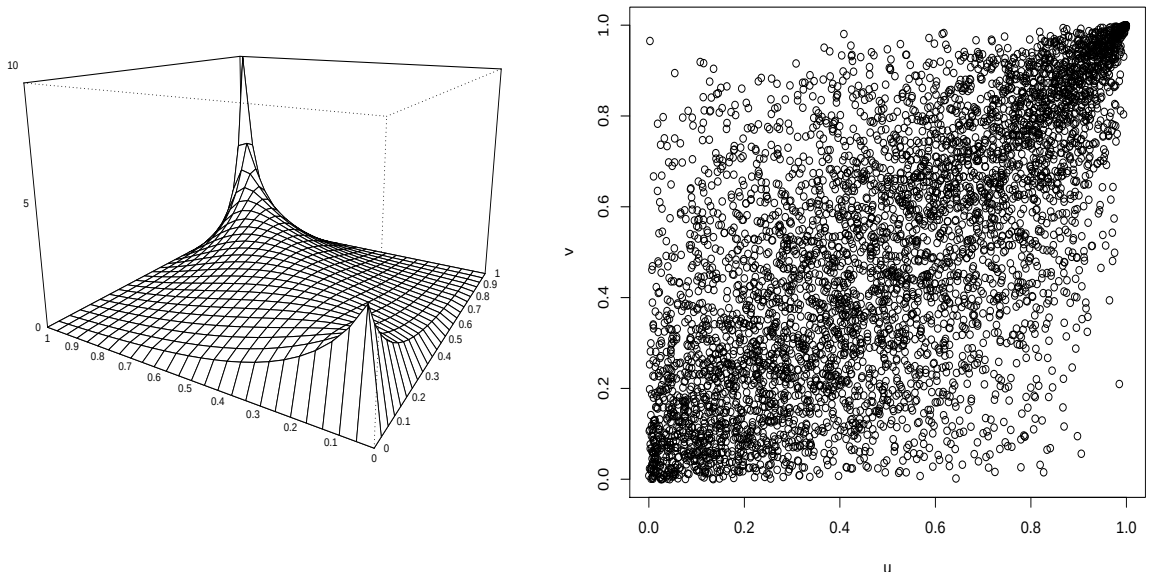

Figure 3.2: Gumbel Copula Model with $\tau = 0.5$

For this copula, the dependence parameter and Kendall's τ are linked via $\theta_g = 1 - \tau$. 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. Using this method, data generation therefore requires an iterative solution which is computationally intensive. An R package is available to generate such data, however since the application of the two-stage meta-analytic copula method within this simulation study was conducted using SAS® software, data were instead generated using the mixtures of powers algorithm described by Trivedi and Zimmer (2007), to avoid mixing between different statistical software packages. This algorithm is based on the work of Marshall and Olkin (1988), and describes how mixture distributions can be used to represent copulas and allow data generation from the marginal distributions, thus removing the need for any iterative procedures. New code was therefore created to generate datasets with the dependence structure of the Gumbel copula as well as the underlying trial-level surrogacy.

The algorithm for data generation from a Gumbel copula can also be simplified into three steps:

Step 1:

The first step of the algorithm starts by generating two independent and identically distributed uniform variables from $Un(0,1)$, U_{ij} and V_{ij}, exactly as was done for the Clayton copula. In addition, a random variable, γ, also needs to be generated from a positive stable distribution. In order to generate γ, a number of steps need to be taken:

- Variables η and w are drawn from $U(0,\pi)$ and the standard exponential distribution, respectively.
- A value z is derived using η and the copula parameter θ_g as $z = \frac{\sin(\eta(1-\theta_g))(\sin(\eta\theta_g))^{\frac{\theta_g}{1-\theta_g}}}{\sin(\eta)^{\frac{1}{1-\theta_g}}}$.
- A value of γ is subsequently derived using $\gamma = \left(\frac{z}{w}\right)^{\frac{1-\theta_g}{\theta_g}}$.

Step 2:

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

$$\begin{aligned} S^0_{ij} &= \exp\left(-\left(\frac{-\log(U_{ij})}{\gamma}\right)^{\theta_g}\right), \\ T^0_{ij} &= \exp\left(-\left(\frac{-\log(V_{ij})}{\gamma}\right)^{\theta_g}\right). \end{aligned}$$

For the Gumbel copula, the parameter θ_g can be calculated directly from Kendall's τ using $\theta_g = 1-\tau$, and so values of θ_g of 0.8, 0.5 and 0.2 achieve 'true' individual-level association of 0.2, 0.5 and 0.8 respectively.

Step 3:

S^0_{ij} and T^0_{ij} are transformed to exponentially distributed time-to-event variables, S_{ij} and T_{ij}, using the same method used in Equations 3.2 and 3.3. Again, this provides two time-to-event variables that are associated according to the underlying data structure of the Gumbel copula, with strength of association controlled by the parameter θ_g. Censoring was applied by generating an exponential variable C_{ij} and comparing to S_{ij} and T_{ij}. 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.

3.2.6 Selection of Simulation Parameters

In addition to simulation factors presented in Table 3.1, and the values of θ_c and θ_g defined in Sections 3.2.4 and 3.2.5, appropriate values of the fixed parameters used to transform the uniform variables to time-to-event outcomes, including treatment effects, baseline hazards and the censoring distribution, need to be selected. These parameters define the average time to disease progression and death, as well as the proportion of patients in the simulated clinical trial datasets who remain event-free at the end of observation.

Following consideration of multiple clinical trial datasets as case studies, parameters were selected based on results of two Phase III clinical trials investigating two molecules for the treatment of HER2 positive gastric cancer (Hecht et al., 2016; Bang et al., 2010). These trials were selected as they investigated different molecules with the same intended mechanism of action (targeting the HER2 protein) and provided very similar median PFS and OS times. Whilst the hazard ratios varied slightly between the trials, the observed values represent the general strength of treatment effect that is often planned for new clinical trials for these endpoints. The following values were therefore selected to generate the simulated datasets:

$\lambda_S = 0.18$, baseline hazard function for a median time to surrogate outcome of approximately $5 - 6$ months;

$\lambda_T = 0.07$, baseline hazard function for a median time to true outcome of approximately $10 - 11$ months;

$\alpha = -0.4$ (logarithm of hazard ratio), corresponding to a hazard ratio for S of approximately 0.67;

$\beta = -0.2$ (logarithm of hazard ratio), corresponding to a hazard ratio for T of approximately 0.82;

$\lambda_C = 0$, 0.027 and 0.1, used to derive censoring proportions of 0%, 30% and 60% by generating a value from a standard exponential distribution and dividing by λ_C.

Use of these simulation parameters allowed for a number of realistic scenarios that are commonly encountered in the analysis of clinical trial data. First, the time to observation of the surrogate endpoint was considered to be approximately half of that for the true endpoint, representing scenarios where surrogates would be considered worthy of use. Further, the treatment effect on the surrogate is assumed to be slightly stronger than that on the true endpoint, since such a relationship is commonly observed, where overall survival can be confounded by additional follow-on therapies administered to a patient once they have experienced disease progression. A Phase III trial designed with a primary endpoint of OS using these parameter values, with commonly used Type I error of 5% and Type II error of 20%, would lead to a study of approximately 850 patients over a total expected duration of 120 months. Use of the surrogate endpoint would reduce this significantly to approximately 230 patients with a total duration of 36 months. These parameters are therefore considered to reliably reflect a setting where there would be strong interest in evaluating potential surrogate endpoints. Additional assumptions were imposed on the simulated datasets to accurately reflect situations observed in real-life clinical trial data:

- A small proportion of patients who 'died' were considered censored for PFS but an event for OS, to reflect patients who initiate alternative anti-cancer therapy without evidence of disease progression, or experience an extended period of time prior to death during which disease assessments are not performed ($\approx 5\%$ using a Bernoulli distribution).

- When the generated time for OS was censored, and the generated time of TTP or PFS was shorter (i.e. $S < T$), the surrogate was considered as an event 80% of

the time. This allows approximately 20% of subjects to be censored for TTP/PFS earlier than OS, representing scenarios where subjects withdraw consent from further invasive medical procedures to determine disease status, or to allow for the time-lag between disease assessments (i.e. last known alive date may be later than the last known disease state).

A total of 5,000 repetitions for each scenario were run, with simulation and analysis conducted on a Windows 7 64-bit machine with 4GB RAM, using macros based on SAS® software, Version 9 for Windows. Across all scenarios, copula model parameters, including the dependence parameter (and therefore the individual-level surrogacy) and the trial-specific treatment effects were estimated based on the maximum likelihood approach using a Newton-Raphson procedure. This was implemented using SAS® software procedure NLPNRR (Newton-Raphson ridge optimisation method).

3.3 Results

Results of the simulation study are presented for each of the factors described in Section 3.2. In order to improve readability of the large number of scenarios, the results are presented first for the surrogate endpoint of time-to-progression in Section 3.3.1, followed by progression-free survival in Section 3.3.2. Within each of these sections, estimation of individual-level surrogacy and trial-level surrogacy are presented separately. Due to similarity of results between the parameter controlling the ranges of treatment effects across trials ($\sigma = 0.1, 0.2$), only the results for the smaller ranges are presented herein; remaining results can be found in Appendix A (Figures A.1 to A.8) for both individual and trial-level surrogacy.

Within each section, the convergence of the two-stage meta-analytic copula method is first discussed. Following this, a review in the performance of the method in estimating τ and R^2_{trial} is provided. Results across all 5,000 simulated datasets are presented in the form of boxplots. Each scenario is presented using a figure containing nine individual plots,

showing all combinations of the individual and trial-level surrogacy; each row contains a fixed individual-level value, and each column a fixed trial-level value. Within each of these individual plots are results for the fixed combination of τ and R^2_{trial} across all numbers of trials (N=4, 6), patients within each trial (n=80, 120, mixed) and proportions of censoring (0%, 30%, 60%). Labels above each plot indicate which combination of surrogacy values is being displayed, and labels underneath the overall figure detail the scenario being presented. Estimates considered to be outliers are not presented (values are considered outliers if they lie below the first quartile or above the third quartile by a margin of 1.5 times the inter-quartile range). To support the graphical displays, summary tables are included to show the median percentage bias across all simulation runs (calculated as the percentage difference between the estimated value of τ or R^2_{trial} and the reference value, as a proportion of the reference value). To improve readability, only the largest sample sizes are included in these summary tables ($N = 6$, $n = 120$).

3.3.1 Time-to-Progression

Convergence

When using TTP as the surrogate endpoint, there were very few issues of non-convergence, with a maximum of 56/5000 runs (1.12%) across all scenarios investigated. The majority of this non-convergence was for low true individual-level association, however the very low number of simulation runs with convergence problems indicates that this issue is not of concern when considering TTP as the surrogate endpoint.

Estimation of τ

Figure 3.3 presents estimated values of τ when using TTP as the surrogate endpoint and a Clayton copula model to generate the data. As noted previously, this data generation algorithm is considered to be the ideal case where the assumptions of the model are

'correct', such that there is no model misspecification. In this scenario, it is therefore expected that the performance of the two-stage meta-analytic copula method would be reasonable, subject to the small sample sizes used in estimation of model parameters.

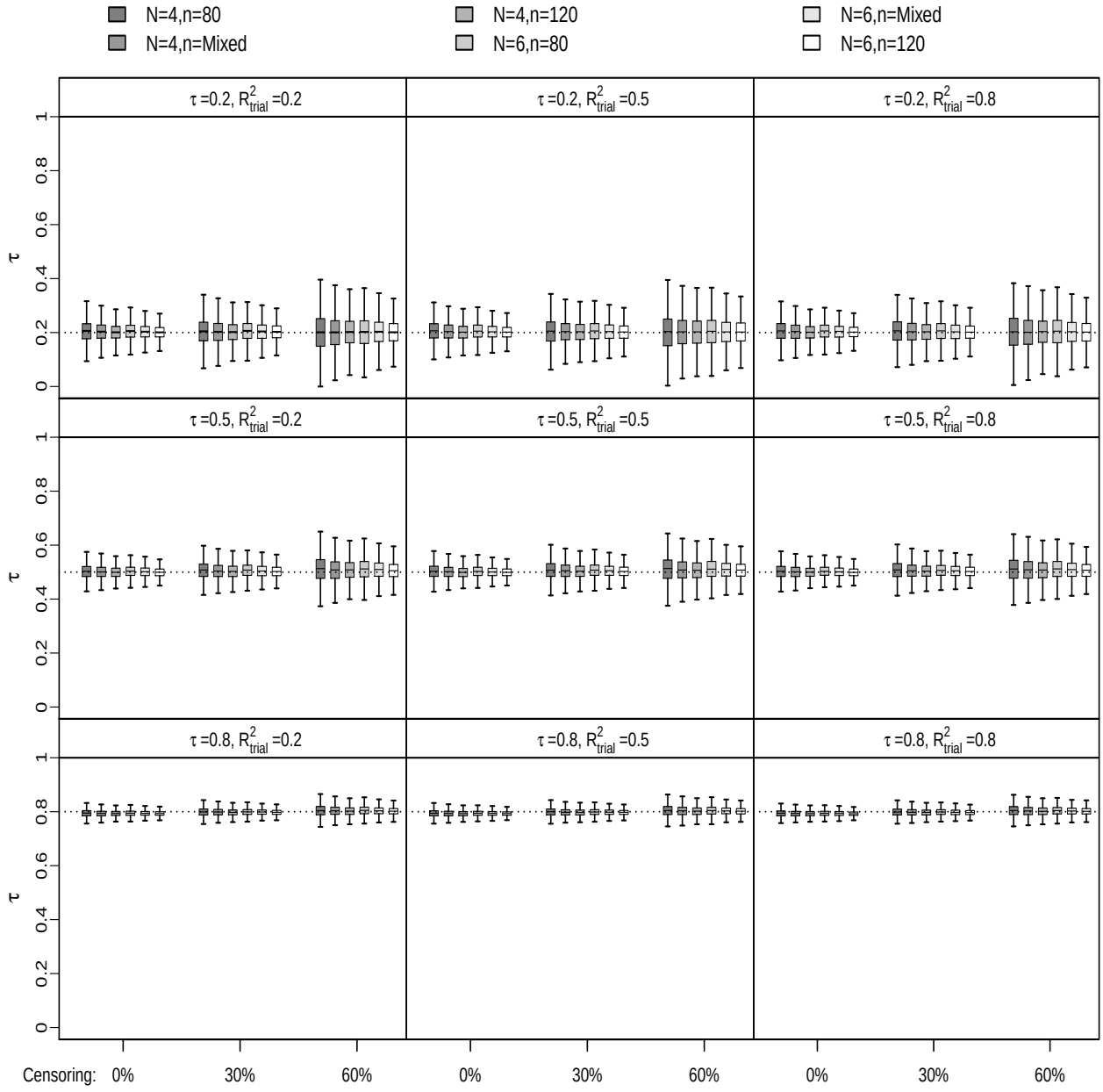

Figure 3.3: Boxplots of estimates of τ: TTP, Clayton Copula Data Generation, Clayton Copula Application

Results of the simulations demonstrate that estimation of τ is generally good, with estimates being close to the reference value and variability increasing only slightly with censoring. For high levels of individual association, there is very little variability in estimates, even for the smallest sample sizes, and identification of a strong surrogate endpoint is therefore possible with high reliability. However, as the true level of individual association decreases, the variability increases, with high variability observed in results for true $\tau = 0.2$. That said, in these scenarios the estimated τ values do not exceed 0.4 even under the highest level of censoring, a value which could be considered as sufficiently low to conclude that surrogacy is poor. The higher variability does not therefore prevent a reliable conclusion. For the medium level of individual association ($\tau = 0.5$), the range of estimates is approximately 0.4 to 0.65, reflecting moderate strength of association and reasonably representing the input value. Across all scenarios, variability appears to reduce slightly with increased sample size and number of trials, as expected, however the performance of the method appears generally good even when limited data are available. This finding is supported by the summary of percentage bias presented in Table 3.2, where the (median) absolute bias of τ never exceeds 1.5%. Whilst not of primary interest here, confidence intervals for τ based on the example scenario of Clayton copula generated datasets with $R^2_{trial} = 0.5$, $N = 4$ and $n = 80$ are presented in Appendix Figure A.11 to illustrate the certainty in the estimates of τ.

Table 3.2: % Bias of Estimates of τ and R^2_{trial}: $N = 6$, $n = 120$, TTP with Clayton Data

			Median % bias	
τ	R^2_{trial}	% Censoring	τ	R^2_{trial}
0.2	0.2	0%	0.413	22.745
0.2	0.5	0%	0.859	-14.362
0.2	0.8	0%	0.787	-24.833
0.2	0.2	30%	1.303	12.427
0.2	0.5	30%	0.984	-24.244
0.2	0.8	30%	0.717	-33.441
0.2	0.2	60%	0.491	-4.296
0.2	0.5	60%	0.951	-41.116
0.2	0.8	60%	0.803	-50.992
0.5	0.2	0%	-0.111	62.096
0.5	0.5	0%	-0.153	16.741
0.5	0.8	0%	-0.212	-0.293
0.5	0.2	30%	0.424	64.362
0.5	0.5	30%	0.472	11.617
0.5	0.8	30%	0.484	-4.880
0.5	0.2	60%	1.046	57.105
0.5	0.5	60%	1.515	-3.911
0.5	0.8	60%	1.422	-22.402
0.8	0.2	0%	-0.834	72.064
0.8	0.5	0%	-0.852	22.395
0.8	0.8	0%	-0.937	7.298
0.8	0.2	30%	-0.300	88.767
0.8	0.5	30%	-0.286	28.222
0.8	0.8	30%	-0.382	7.102
0.8	0.2	60%	0.250	111.625
0.8	0.5	60%	0.256	28.112
0.8	0.8	60%	0.195	2.467

In order to examine the impact of model misspecification, estimates of τ from the data generated using a Gumbel copula are presented in Figure 3.4, with supportive data provided in Table 3.3.

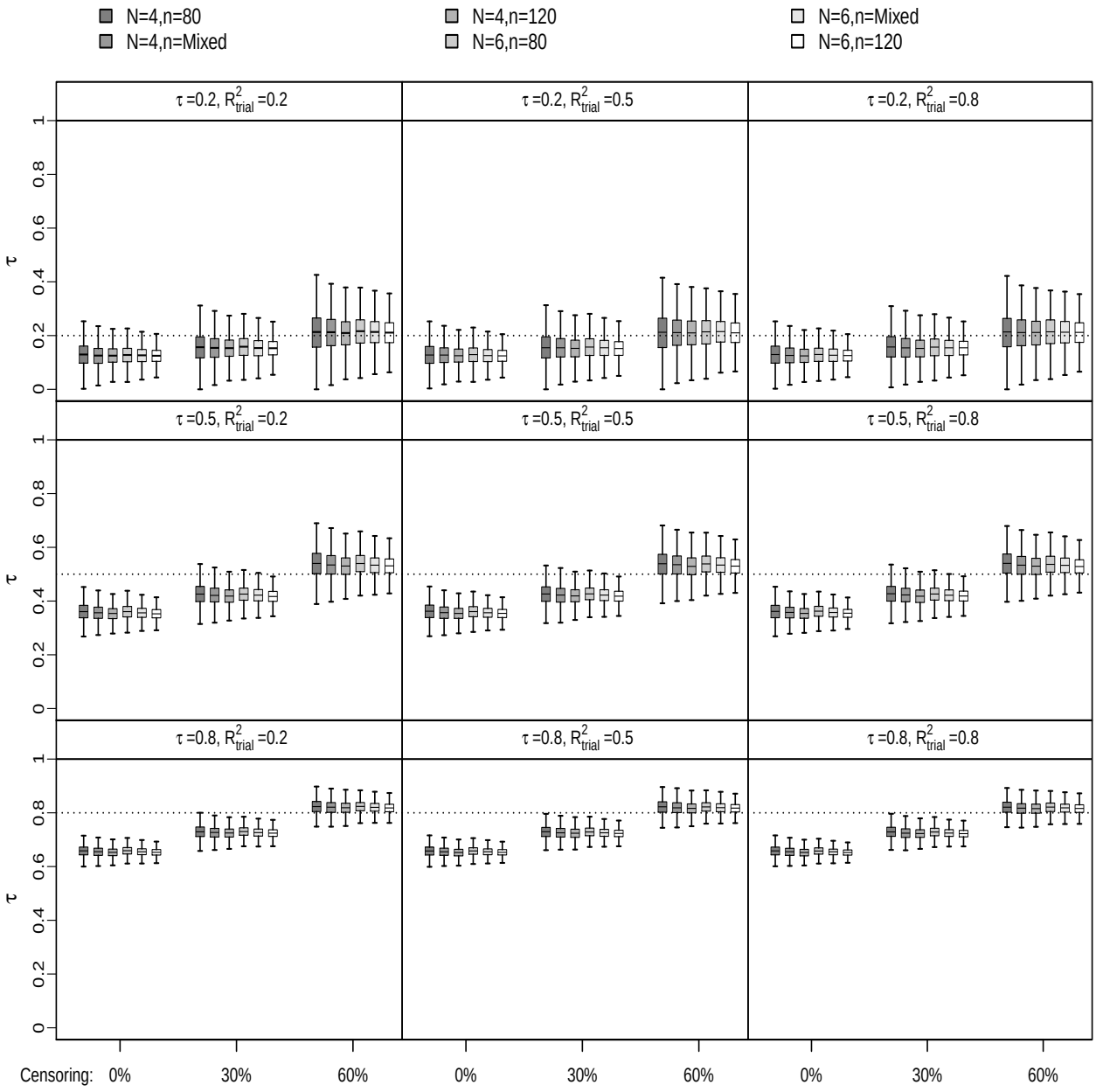

Figure 3.4: Boxplots of estimates of τ: TTP, Gumbel Copula Data Generation, Clayton Copula Application

Table 3.3: % Bias of Estimates of τ and R^2_{trial}: $N = 6$, $n = 120$, TTP with Gumbel Data

			Median % bias	
τ	R^2_{trial}	% Censoring	τ	R^2_{trial}
0.2	0.2	0%	-37.568	4.231
0.2	0.5	0%	-37.752	-26.393
0.2	0.8	0%	-37.197	-34.250
0.2	0.2	30%	-23.508	4.649
0.2	0.5	30%	-24.112	-30.855
0.2	0.8	30%	-22.758	-38.200
0.2	0.2	60%	5.505	0.071
0.2	0.5	60%	5.113	-41.681
0.2	0.8	60%	5.971	-48.630
0.5	0.2	0%	-29.488	45.069
0.5	0.5	0%	-29.152	0.870
0.5	0.8	0%	-29.052	-13.298
0.5	0.2	30%	-16.593	46.099
0.5	0.5	30%	-16.353	-2.023
0.5	0.8	30%	-16.241	-19.336
0.5	0.2	60%	6.212	47.851
0.5	0.5	60%	6.008	-7.867
0.5	0.8	60%	5.709	-27.353
0.8	0.2	0%	-18.374	95.934
0.8	0.5	0%	-18.388	29.514
0.8	0.8	0%	-18.462	3.736
0.8	0.2	30%	-9.448	93.187
0.8	0.5	30%	-9.571	23.929
0.8	0.8	30%	-9.688	0.574
0.8	0.2	60%	2.261	102.216
0.8	0.5	60%	2.182	21.023
0.8	0.8	60%	2.021	-2.674

Given the very different dependence structures assumed by the Clayton and Gumbel copula functions, it is expected that estimation of τ would be poorer when based on the Gumbel data generation, and the results show that this is the case, with two notable changes.

Firstly, regardless of the underlying strength of association, the method appears to provide higher estimates as the percentage of censoring increases. Based on the increasing estimates, it is considered likely that higher censoring would lead to slight over-estimation of the true underlying τ, hence the method could be considered to be reasonably accurate when approximately 50% of patients are censored, but not with lower (causing under-estimation) or higher (likely causing over-estimation) proportions of censoring. There appears to be only minor improvement through increasing the number of trials or sample sizes within trials. Data in Table 3.3 also support these findings, with median percentage bias reaching as low as -38% under no censoring, and $+6\%$ for the highest level of censoring. This issue is discussed further in Section 3.4.

Secondly, the level of variability in the estimated values of τ has increased as compared to the Clayton generated data, and this is apparent across all levels of association. Estimates of truly low levels of association appear to remain low across all scenarios, noting that higher values may be observed when the level of censoring increases above 60%. However, the increased variability in the results for $\tau = 0.5$ means that, particularly when the level of censoring is low, there is overlap in results between truly low and medium levels of surrogacy. Further, whilst the highest levels of association remain with the lowest variability, the medium level of association now provides estimates of individual-level surrogacy that reach as high as 0.7, in some cases with overlap with the lower tails of estimates for truly high surrogacy. This increase in estimates could potentially lead to mediocre surrogates being considered to have high levels of individual association. These results therefore demonstrate that whilst low ($\tau = 0.2$) and high ($\tau = 0.8$) surrogacy can be reliably identified, medium levels of association are less clear, and with small sample sizes could be misleadingly interpreted as being either too weak or too strong. The im-

portance of verifying the underlying dependence structure of the data, and the adequacy of model fit of the selected copula, is therefore apparent. There is marginal improvement with increases in sample size and numbers of trials, but it is considered that substantially more data would be required to improve estimation.

Overall, the two-stage meta-analytic copula method for assessing TTP as a surrogate endpoint at the individual-level has demonstrated good performance under correctly specified models, with additional variability in the incorrectly specified model that can lead to under- or over-estimation, particularly when sample sizes are small.

Estimation of R^2_{trial}

Estimates of R^2_{trial} are provided in Figures 3.5 (for Clayton generated data) and 3.6 (for Gumbel generated data). In these figures, the left column shows $R^2_{trial} = 0.2$, middle column $R^2_{trial} = 0.5$ and right column $R^2_{trial} = 0.8$, with the rows showing the impact of increasing the individual-level association. Since the results appear very similar between the two data generation algorithms, description of the results will not distinguish between the two. Supportive data are again included in Tables 3.2 and 3.3 for the largest sample sizes.

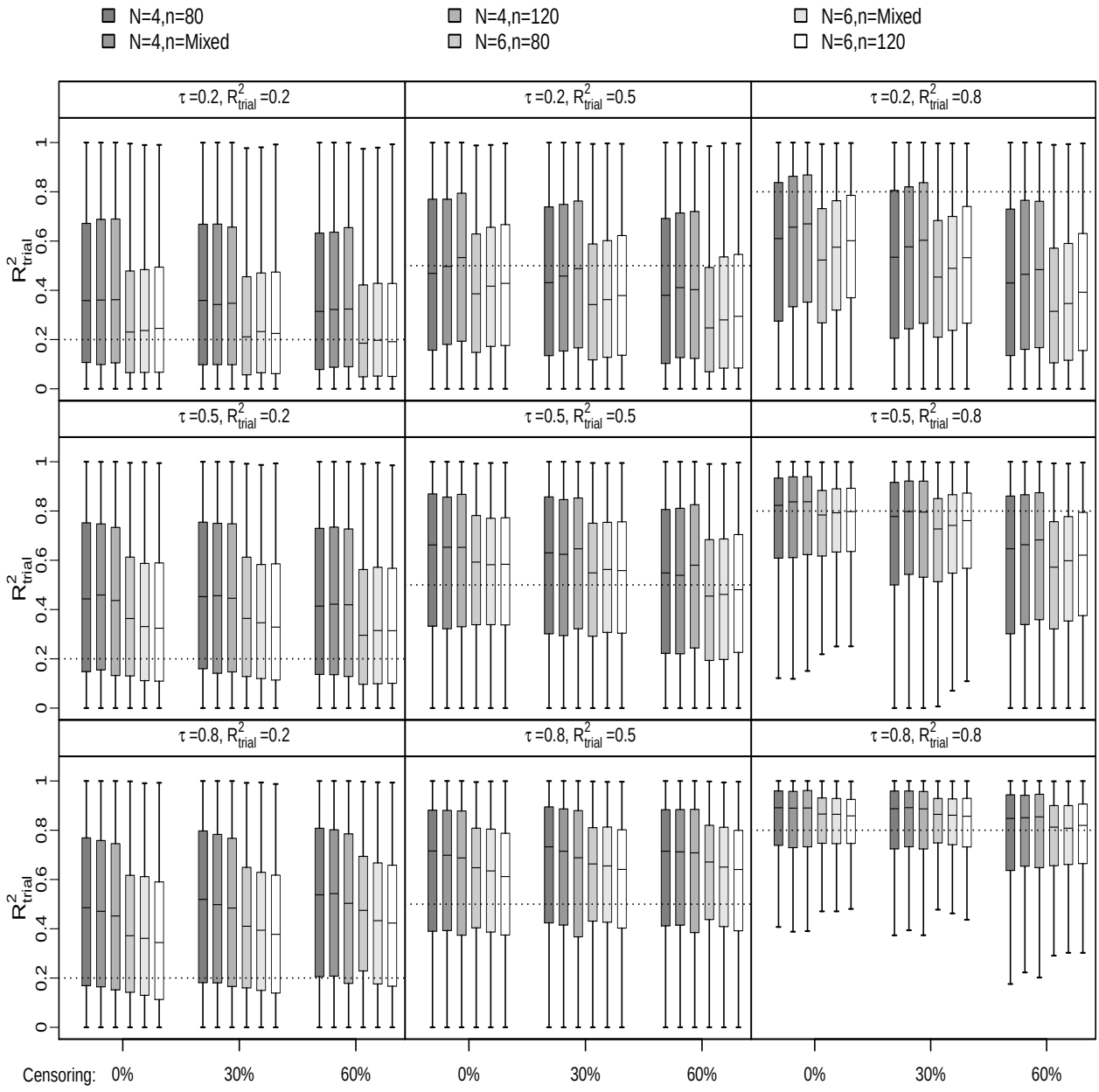

Figure 3.5: Boxplots of estimates of R^2_{trial}: TTP, Clayton Copula Data Generation, Clayton Copula Application

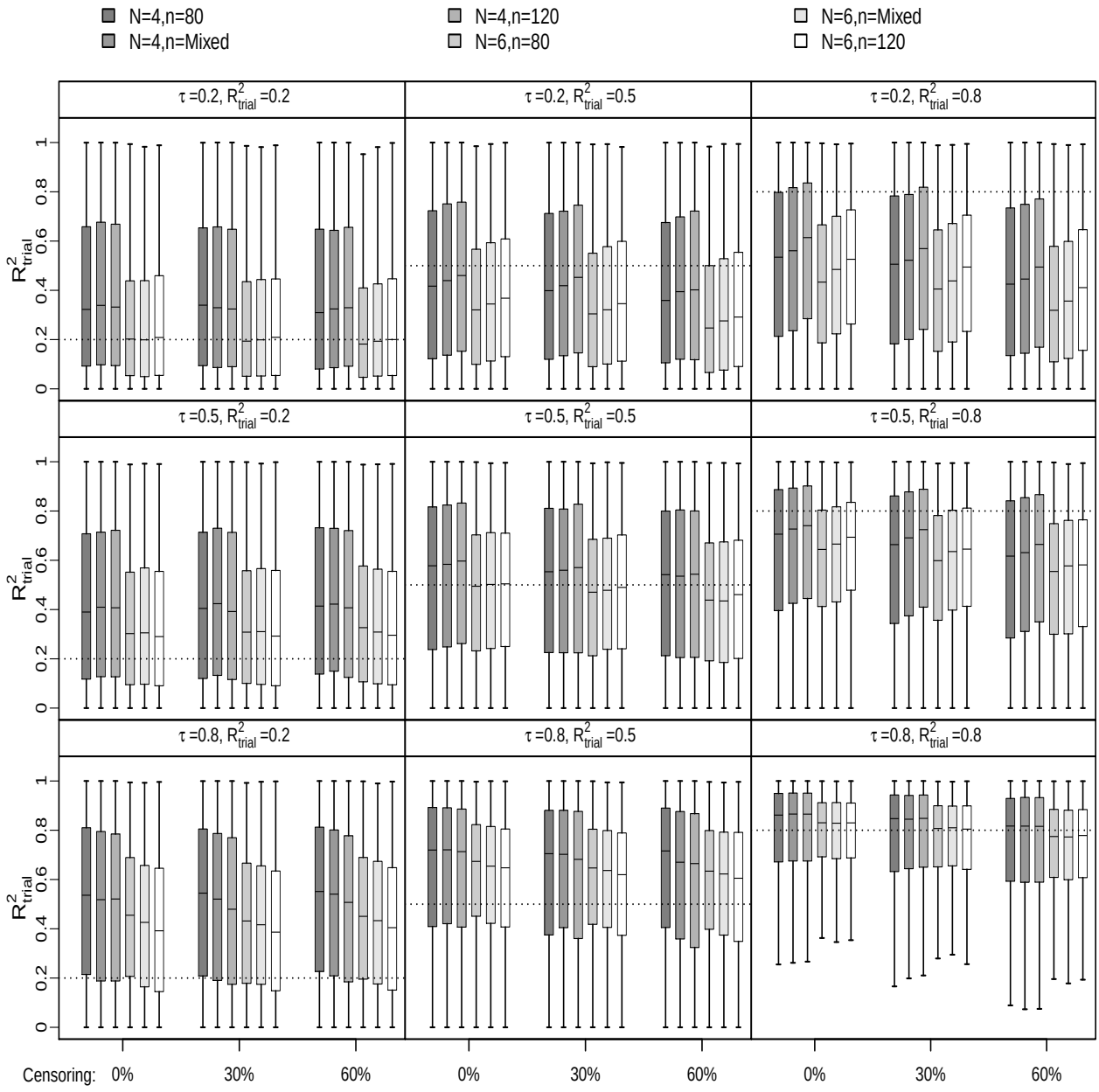

Figure 3.6: Boxplots of estimates of R^2_{trial}: TTP, Gumbel Copula Data Generation, Clayton Copula Application

As might be expected from the small number of trials, estimation of R^2_{trial} is overall poor, with severe under- or over-estimation of the true value and very high variability. Although the median results are sometimes close to the true value, the estimates lie across the entire unit interval in the majority of scenarios. This is considered to be a result of the

small number of trials used in the simulation. Although the estimates appear to increase slightly with increased underlying association, they are far from the true (reference) values and do not provide reliable interpretation.

There are two notable features of the results; the impact of increasing the number of trials, and the dependency between R^2_{trial} and τ. It is evident that the increase from 4 to 6 trials has reduced the variability, with a smaller interquartile range as depicted by the grey shaded boxes, although the impact decreases with higher R^2_{trial} values. Despite this, the range of estimates remains across the unit interval, and it can be reasonably concluded that estimation of R^2_{trial} cannot be reliably achieved with such low numbers of trials. Splitting the trials into smaller subsets would increase the number of units available for analysis, however this is not recommended for these small sample settings since each study has only $80-120$ patients each; splitting these into multiple subgroups would likely lead to increased bias in estimation of R^2_{trial}, as was observed by Renfro et al. (2014).

With regards to dependency between R^2_{trial} and τ, it can be seen from the right column ($R^2_{trial} = 0.8$) that estimates are increasing in value across the rows of the figure, showing an increase in estimates of R^2_{trial} with increasing τ. The effect is less pronounced for the low and middle values of R^2_{trial}, but appears to remain present. This finding was also observed by Burzykowski (2001) and is a result of the use of the estimated treatment effects from stage one of the analysis without correcting for the estimation error. When this is not corrected for, the correlation between estimated treatment effects can under- or over-estimate the true correlation between the treatment effects depending on the size of correlation between the measurement errors. Whilst the use of adjusted estimators would be preferred, such approaches are currently considered to be difficult to use in practice due to issues of non-convergence, as noted in Section 2.3.5. Overall, results have demonstrated that for the scenarios under investigation here, the estimation of R^2_{trial} cannot be reliably performed using the two-stage meta-analytic copula method.

3.3.2 Progression-Free Survival

Convergence

The change in surrogate endpoint from TTP to PFS has a notable effect on the non-convergence of the two-stage meta-analytic copula method, which increases to a maximum of 61.3% (3063/5000 runs). However, such high rates of non-convergence occur exclusively for scenarios based on low individual-level surrogacy. For medium-high individual-level surrogacy, the non-convergence remains at a rate of zero for the majority of scenarios, reaching a maximum of 1.06%. The complex nature of the joint modelling required by the two-stage meta-analytic copula method therefore appears to cause significant problems when used with an endpoint that does not satisfy the symmetry assumption of copula models.

Estimation of τ

Based on outcomes simulated using the Clayton copula, the estimates of τ from the application of PFS as the surrogate endpoint are presented in Figure 3.7, with supportive information in Table 3.4. Given the correct model specification in terms of the dependence structure, the results provide a direct assessment of the impact from incorrectly assuming symmetry of surrogate and true endpoints.

These plots of individual estimates of τ highlight several aspects worthy of consideration. First, it is apparent that even under correctly specified models, the method substantially over-estimates low individual-level association across all 5,000 simulation runs, with over-estimation of moderate level association also being observed. Further, in both of these settings, there is a notable worsening caused by increased censoring within the datasets, with little overlap between estimates of τ between the datasets with no censoring and those with high (60%) censoring. Under this highest level of censoring, a true level of association of 0.2 could be estimated as high as 0.7 (median bias approximately 150%),

which would almost certainly be considered encouraging enough to move forward with a surrogate. This over-estimation is considered a key finding, since many of the applications of the two-stage meta-analytic copula approach in practice have evaluated the surrogacy of PFS for OS. Indeed, many clinical trials that have already achieved regulatory approval based on an alternative endpoint to overall survival have done so through use of PFS, and this endpoint remains the first choice for many settings where OS is not feasible. These results therefore demonstrate that this could be a major issue in practice. Interestingly, the truly high level of association continues to be estimated with high reliability and precision, reflected by a median bias of $< 5\%$ across all scenarios with the largest sample sizes, and estimates of τ ranging from approximately 0.75 to 0.92.

Whilst the variability in estimates is high, particularly for low levels of association, it is reasonably similar to that observed when TTP was used as the surrogate. Consistent with that setting, the variability remains low for the highest level of association, and increases as the true τ decreases. Variability is also higher for censored data. However, when PFS is used as the surrogate, there is greater improvement in the variability of estimates by increasing the number of trials and the sample sizes. The range of estimates of τ appears to be half for the largest sample sizes ($N = 6$, $n = 120$) as compared to the smallest sample sizes ($N = 4$, $n = 80$). This improvement is greater than was observed for TTP.

Given this deterioration in performance under correct model specification, it could be expected that the results would be impacted further when the model being used does not follow the underlying structure of the data, and results of the Gumbel data generation confirm that this is the case (Figure 3.8, Table 3.5).

From Figure 3.8, it can be seen that for low levels of association, the two-stage meta-analytic copula method continues to over-estimate quite substantially the true τ, with estimates reaching as high as 0.69 (median bias for largest sample sizes of approximately 150%), and not lower than the reference value of 0.2. Whilst the absolute value of the estimates has decreased slightly for low-medium levels of association as compared to the Clayton data, the severe over-estimation under high censoring remains.

Table 3.4: % Bias of Estimates of τ and R^2_{trial}: $N = 6$, $n = 120$, PFS with Clayton Data

			Median % bias	
τ	R^2_{trial}	% Censoring	τ	R^2_{trial}
0.2	0.2	0%	53.960	132.166
0.2	0.5	0%	52.052	23.194
0.2	0.8	0%	51.533	-7.494
0.2	0.2	30%	105.392	122.543
0.2	0.5	30%	103.747	20.894
0.2	0.8	30%	102.839	-15.145
0.2	0.2	60%	152.067	113.914
0.2	0.5	60%	149.861	4.969
0.2	0.8	60%	147.605	-28.031
0.5	0.2	0%	11.776	119.562
0.5	0.5	0%	11.296	30.187
0.5	0.8	0%	10.634	4.140
0.5	0.2	30%	18.391	125.809
0.5	0.5	30%	17.750	27.614
0.5	0.8	30%	17.118	-1.369
0.5	0.2	60%	29.994	120.573
0.5	0.5	60%	29.735	15.281
0.5	0.8	60%	29.112	-13.597
0.8	0.2	0%	1.696	110.874
0.8	0.5	0%	1.406	26.389
0.8	0.8	0%	1.177	7.596
0.8	0.2	30%	2.913	126.163
0.8	0.5	30%	2.565	31.577
0.8	0.8	30%	2.351	6.838
0.8	0.2	60%	4.838	128.277
0.8	0.5	60%	4.652	29.855
0.8	0.8	60%	4.375	1.915

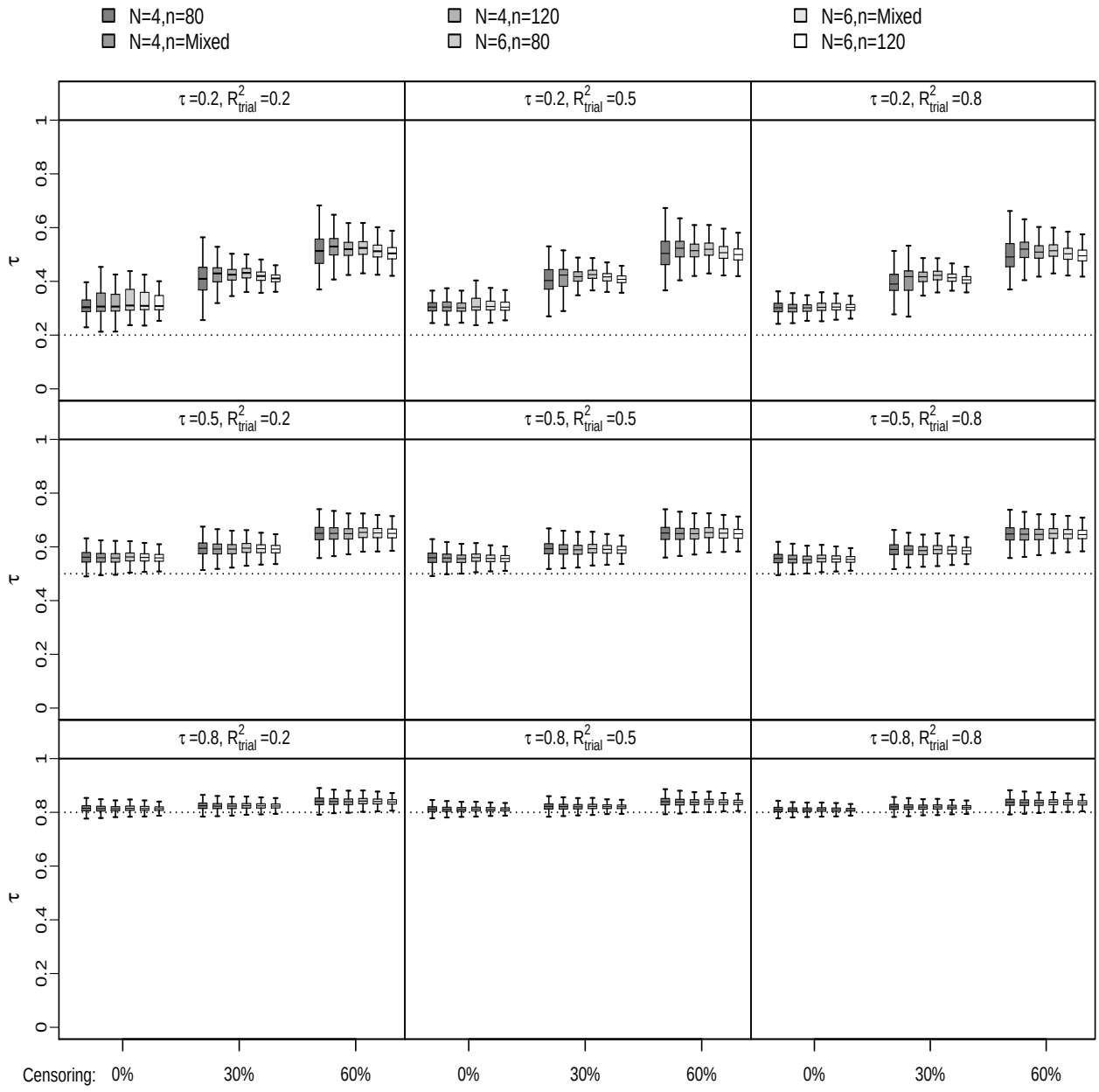

Figure 3.7: Boxplots of estimates of τ: PFS, Clayton Copula Data Generation, Clayton Copula Application

For medium individual-level association, results for the highest level of censoring also remain similar, whereas estimates for the low and no censoring scenarios have reduced, again reflecting the larger spread of results across the censoring proportions when using the Gumbel generated data. Variability in the estimates of τ has also increased, with ranges

of approximately 0.35 right up to very high values of 0.7, again highlighting the risk of incorrectly declaring PFS to have strong predictive power for OS. As compared to TTP, all estimates of true $\tau = 0.5$ appear to be slightly higher, leading to reasonable estimation under 30% censoring, but under-estimation under no censoring and over-estimation under censoring $> 60\%$. The impact of censoring between the two potential surrogates is therefore consistent, with higher proportions of censoring leading to higher estimates of τ, which could potentially lead to incorrect conclusions.

Encouragingly, for the highest level of true association between PFS and OS, the method appears to perform reasonably well, with the lowest variability and lowest spread of results between censoring proportions. This is consistent with results from TTP, and in fact there is little difference between the endpoints in this scenario. That said, the increased variability and spread in estimates of τ across all true levels of association suggest that any true underlying association strength could be estimated to be very high. Overall, the variability in results for PFS limits interpretability. When the true association is very strong, the method appears to continue to perform well when the model is specified correctly, but any deterioration from this specification appears to cause issues. There is significant overlap in the estimates of τ across scenarios, whereby mediocre or even poor surrogates could incorrectly be concluded as having strong predictive ability. This is of great concern, since this could lead to such endpoints being used in confirmatory clinical trials.

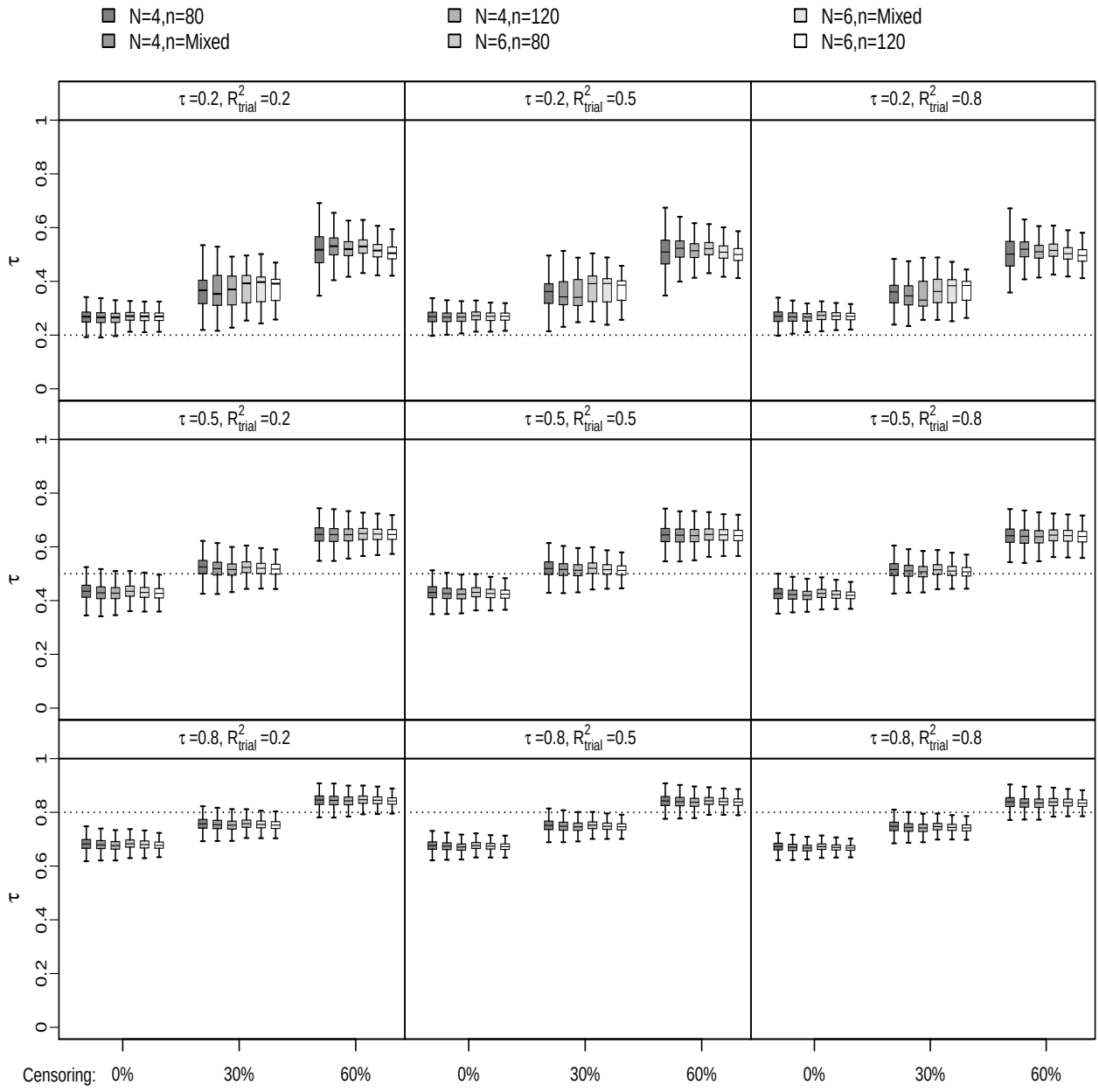

Figure 3.8: Boxplots of estimates of τ: PFS, Gumbel Copula Data Generation, Clayton Copula Application

Table 3.5: % Bias of Estimates of τ and R^2_{trial}: $N = 6$, $n = 120$, PFS with Gumbel Data

			Median % bias	
τ	R^2_{trial}	% Censoring	τ	R^2_{trial}
0.2	0.2	0%	34.734	127.501
0.2	0.5	0%	34.969	18.919
0.2	0.8	0%	34.950	-12.692
0.2	0.2	30%	95.729	149.653
0.2	0.5	30%	92.908	19.584
0.2	0.8	30%	92.372	-18.033
0.2	0.2	60%	152.443	112.511
0.2	0.5	60%	149.963	2.270
0.2	0.8	60%	148.064	-26.445
0.5	0.2	0%	-14.651	143.254
0.5	0.5	0%	-15.245	27.578
0.5	0.8	0%	-16.106	-4.548
0.5	0.2	30%	3.457	136.919
0.5	0.5	30%	2.437	22.696
0.5	0.8	30%	1.381	-10.465
0.5	0.2	60%	29.191	121.701
0.5	0.5	60%	28.365	9.262
0.5	0.8	60%	27.726	-20.645
0.8	0.2	0%	-15.364	152.165
0.8	0.5	0%	-15.989	37.061
0.8	0.8	0%	-16.573	5.359
0.8	0.2	30%	-5.942	140.408
0.8	0.5	30%	-6.740	28.608
0.8	0.8	30%	-7.265	2.202
0.8	0.2	60%	5.256	141.179
0.8	0.5	60%	4.773	26.758
0.8	0.8	60%	4.259	-2.650

Estimation of R^2_{trial}

As for the TTP setting, the performance of the two-stage meta-analytic copula method in estimating R^2_{trial} is very similar for both Clayton and Gumbel generated data, and so discussion of the results for the individual data generation structures will not be separated. The estimated R^2_{trial} values for the Clayton and Gumbel datasets can be found in Figures 3.9 and 3.10 respectively, with supportive data in Tables 3.4 and 3.5.

Overall, results are broadly consistent with those from data generated using TTP as the surrogate endpoint. Whilst estimates of R^2_{trial} appear very slightly higher across the majority of scenarios (with the exception of $\tau = 0.5$, $R^2_{trial} = 0.8$ and $\tau = 0.8$, $R^2_{trial} = 0.8$ where results are of a similar magnitude), the range of estimates stretches across the entire unit interval. Whilst Burzykowski et al. (2005) consider $100 - 200$ patients per trial to be sufficient to use R^2_{trial} to generate "reasonable" results, the current simulation study demonstrates that this is likely only possible when there are a larger number of trials containing this number of patients. Even in the largest sample sizes investigated here ($N = 6$, $n = 120$), estimation of R^2_{trial} is poor, with median bias ranging from approximately -30% to approximately 150%. As shown in Figures 3.9 and 3.10, this is a slight improvement over the scenarios with only four trials included, but not sufficient to consider the method to provide reliable results. There was also no impact from increasing sample size within trials, and no difference among datasets with different proportions of censoring. Finally, the dependency between R^2_{trial} and τ that was observed for the TTP is also present for the PFS setting.

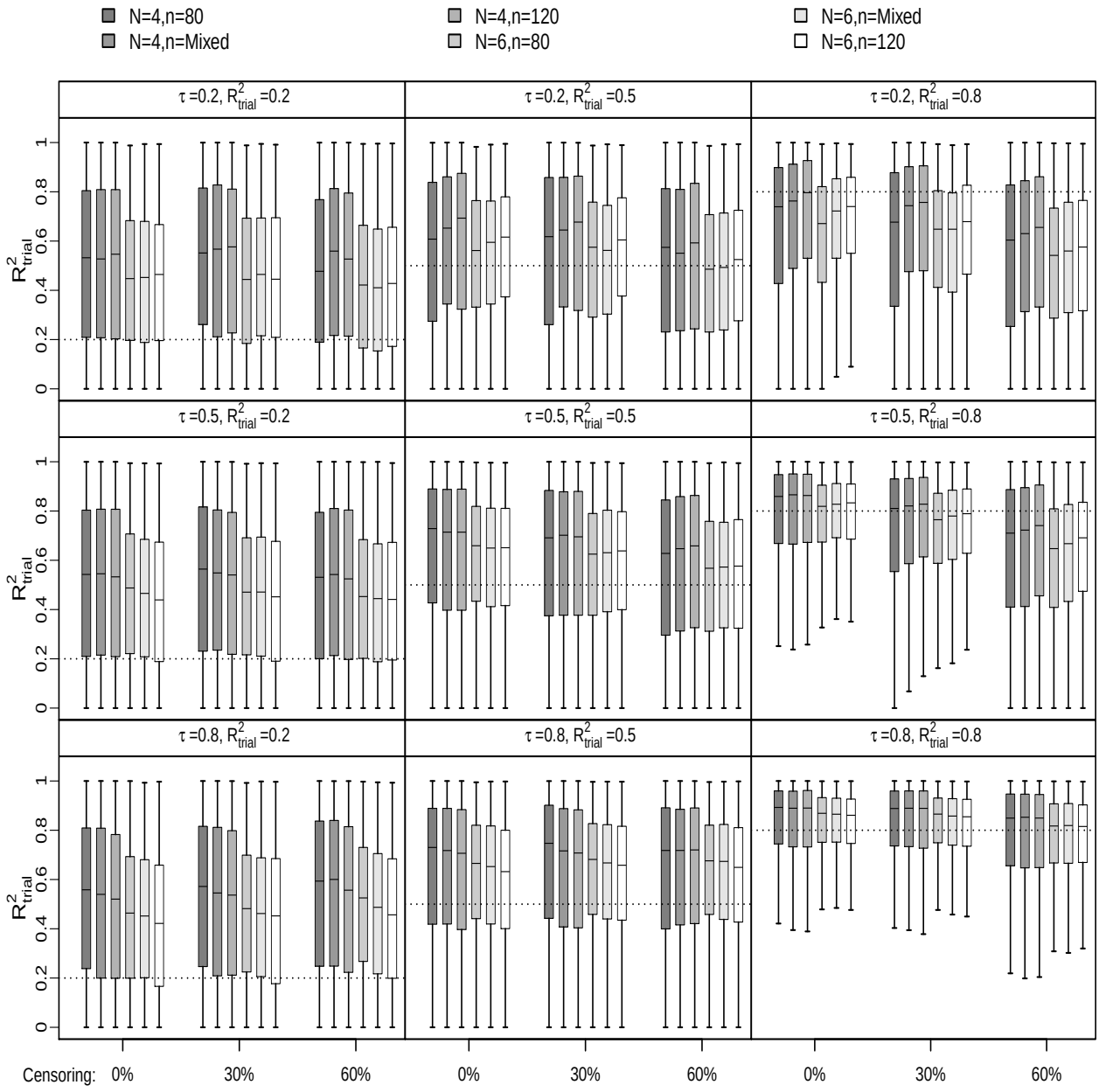

Figure 3.9: Boxplots of estimates of R^2_{trial}: PFS, Clayton Copula Data Generation, Clayton Copula Application

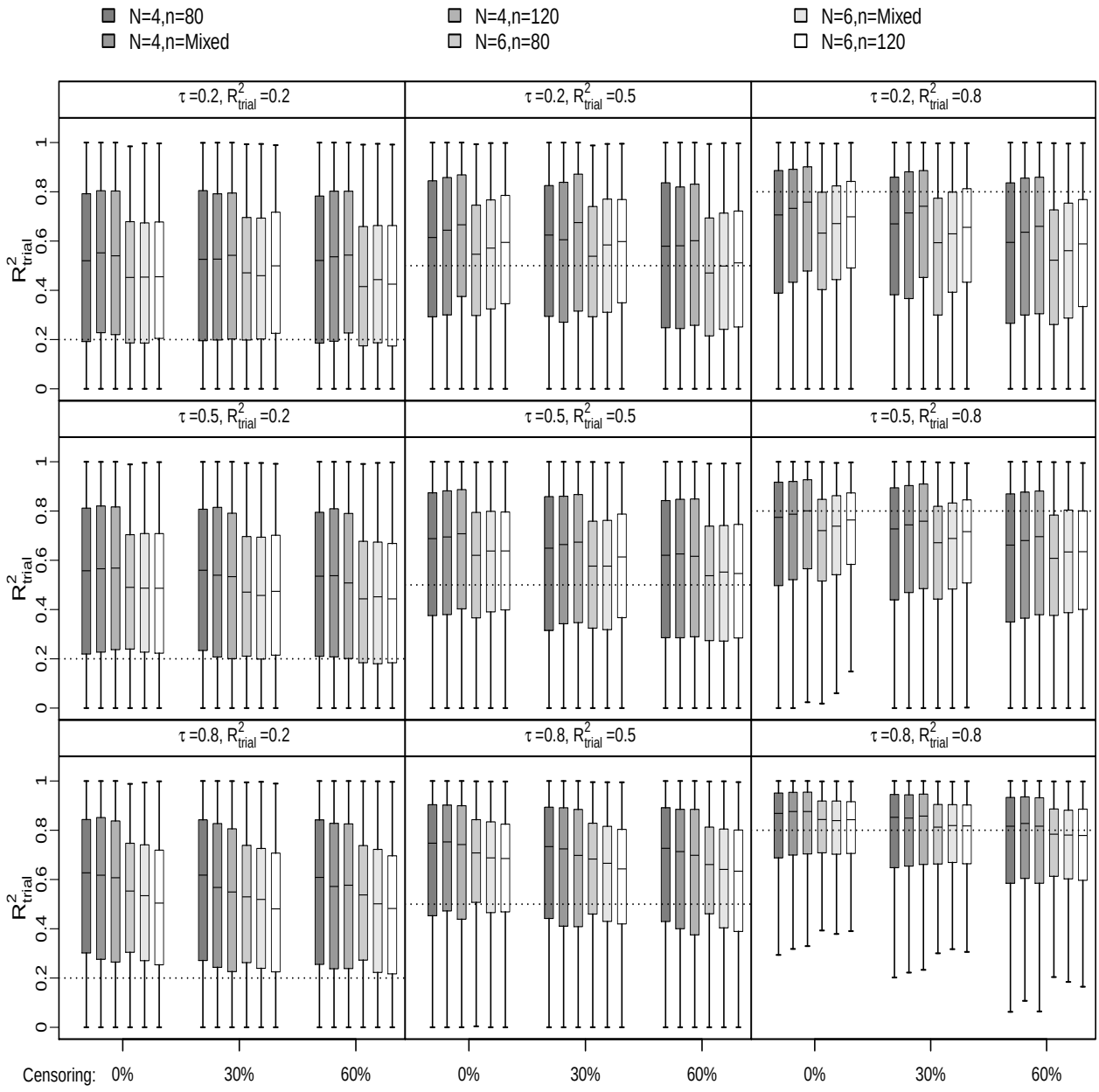

Figure 3.10: Boxplots of estimates of R^2_{trial}: PFS, Gumbel Copula Data Generation, Clayton Copula Application

Summary of Results

The extensive results from all simulation scenarios can be summarised by:

- The two-stage meta-analytic copula method generally performs well in estimating τ when TTP is used as the surrogate endpoint and the correct model specification is assumed. Variability in results increases as τ decreases, but this does not impact interpretation.

- For TTP, changing the dependence structure in the data led to a greater impact of censoring and deteriorated performance, but truly high and truly low surrogacy could be reliably identified. Mediocre surrogates may incorrectly be concluded to be poor or strong predictors for OS.

- When considering PFS as the surrogate endpoint, the method suffers from severe convergence issues and over-estimation of individual-level surrogacy, particularly when the true level of association is low. This finding is apparent even for correctly specified models.

- Under incorrectly specified models, the deterioration with the use of PFS continues, with severe under- or over-estimation of individual-level surrogacy:

  - Severe over-estimation when $\tau = 0.2$, regardless of the level of censoring.

  - Reasonable estimation when $\tau = 0.5$ and there is a low level of censoring, but under-estimation when there is no censoring, and over-estimation under high levels of censoring.

  - Slight over-estimation when $\tau = 0.8$ and there is high censoring, and under-estimation otherwise.

- Regardless of the surrogate endpoint, the method cannot be considered to provide reliable estimates of R^2_{trial} when only a small number of trials are available.

3.4 Understanding the Results

In this section, notable features of the results are described with the aim to understand their cause. First, comparisons with a previous simulation study of the two-stage meta-analytic copula method will be made, followed by discussion of individual elements of the results that require further investigation.

3.4.1 Comparison to Previous Simulation Study

Estimation of τ

The simulation study of Burzykowski (2001) also considered the performance of the two-stage meta-analytic copula method for TTP data generated using the Clayton copula. Although the range of simulation scenarios was slightly different, comparisons between the study of Burzykowski (2001) and that presented in this chapter can be discussed.

Burzykowski (2001) found that estimates of τ were generally positively biased, with $< 1\%$ bias for true $\tau = 0.9$ and $< 4\%$ bias for true $\tau = 0.5$. This bias decreased with increased patient numbers, with a slight increase in bias for increased censoring when $\tau = 0.9$. These findings are consistent with the current study, where absolute percentage bias for the largest sample size ranged from 0.1% to 1.5% for true $\tau = 0.5$ and 0.2% to 0.9% for true $\tau = 0.8$. Similarly, estimation improved marginally when increasing the number of trials and patients within trials.

Whilst Burzykowski (2001) does not provide graphical representation of results to enable a visual assessment of variability, the standard errors of estimates for τ are provided. These values suggest that the standard error decreases as both the number of trials and sample sizes increase, but rises with censoring, and this appears to be independent of the trial level strength of association. Results presented in Section 3.3.1 demonstrate a consistent pattern, with the increase from 4 to 6 trials and 80 to 120 patients per trial leading to reduced ranges of estimates. The increase in variability due to increased

censoring is also present, although this is much more pronounced for the lower levels of association, a scenario not explored by Burzykowski (2001). Overall, results between the two studies are similar, and suggest that altering the treatment effects and other simulation parameters has had minimal impact on estimation of τ.

Estimation of R^2_{trial}

Burzykowski (2001) examined true trial-level association of 0.5 or 0.9, and concluded that the bias in estimates of R^2_{trial} were dependent on the level of τ due to the level of correlation between the measurement errors in the trial specific treatment effects. Results showed a positive bias in estimates of R^2_{trial} for true $\tau = 0.9$ and $R^2_{trial} = 0.5$, and negative bias when $\tau = 0.5$ and $R^2_{trial} = 0.9$. Results from the current study are broadly consistent in that positive bias was observed when $\tau = 0.8$ and $R^2_{trial} = 0.5$, and negative bias observed when $\tau = 0.5$ and $R^2_{trial} = 0.8$ and there was censoring present within the data.

Whereas no difference was observed when increasing the number of trials from 10 to 20 in the study of Burzykowski (2001), improvement was seen in the current study when increasing from 4 to 6 trials. This difference in conclusions is considered to be as a result of the low number of trials included in the current study, where the addition of just two trials could be expected to improve the estimation. Conversely, the current study showed no improvement in estimation through an increase in sample size per trial from 80 to 120, whereas Burzykowski (2001) demonstrated an improvement when increasing sample size from 50 to 200 per trial. It would seem reasonable to assume that this larger (four-fold) increase in sample size has a more substantial impact on parameter estimation than the more modest increase from 80 to 120. Interestingly, Burzykowski (2001) conclude that sample sizes of $100 - 200$ patients per trial is sufficient to reduce the percentage bias to 10% or lower, which is far lower than the results observed in the current study. This is likely due to the low number of trials included in the current simulations.

3.4.2 Variability and Model Misspecification

With regard to results of the current study, it has been demonstrated that the variability in estimates of τ appears to decrease as the absolute value increases, and this is consistent across both Clayton and Gumbel data generation, and for both surrogate endpoints. This result is considered to be due to the use of the Clayton copula function in the surrogacy evaluation approach. Based on this copula model, the variance of Kendall's τ is calculated as

$$V(\widehat{\tau}) = \frac{4V(\widehat{\theta_c})}{(\widehat{\theta_c}+1)^4},$$

where $\widehat{\theta_c}$ is the estimated value of the copula dependence parameter (Burzykowski, 2001). Therefore, the variance in $\widehat{\tau}$ decreases as the absolute value of $\widehat{\theta_c}$ increases. Since the value of θ_c increases as the strength of association increases, this variability decreases as τ gets larger, likely leading to estimates of τ that are smaller in range.

When investigating the impact of model misspecification, the Clayton copula was applied to data generated using a Gumbel copula. Since the Clayton model assumes strong late-tail dependence, it could be expected that the model under-estimates τ when applied to data that is designed to have weak late-tail dependence, as in the Gumbel data. Hence, lower estimates of τ under 0% censoring based on Gumbel data are not unexpected. Under TTP, the increase in estimates with increasing proportion of censoring appears present only for the Gumbel generated data, suggesting that this also is caused by the inappropriate model assumptions. In this case, it is likely that the longest values of S and T are truncated through the censoring, and these values are likely those that have the weakest association. Elimination of such values from the dataset may therefore lead to the effect of stronger association overall, and subsequent increases in values of estimated τ. This can be seen in the scatterplots presented in Figure 3.11, which contain values of S and T generated for 1,000 patients from the Gumbel copula with $\tau = 0.8$. As the percentage of censoring increases (Figure 3.11(a) to 3.11(c)), the weakly correlated values in the upper tail are removed, and overall association between endpoints appears stronger.

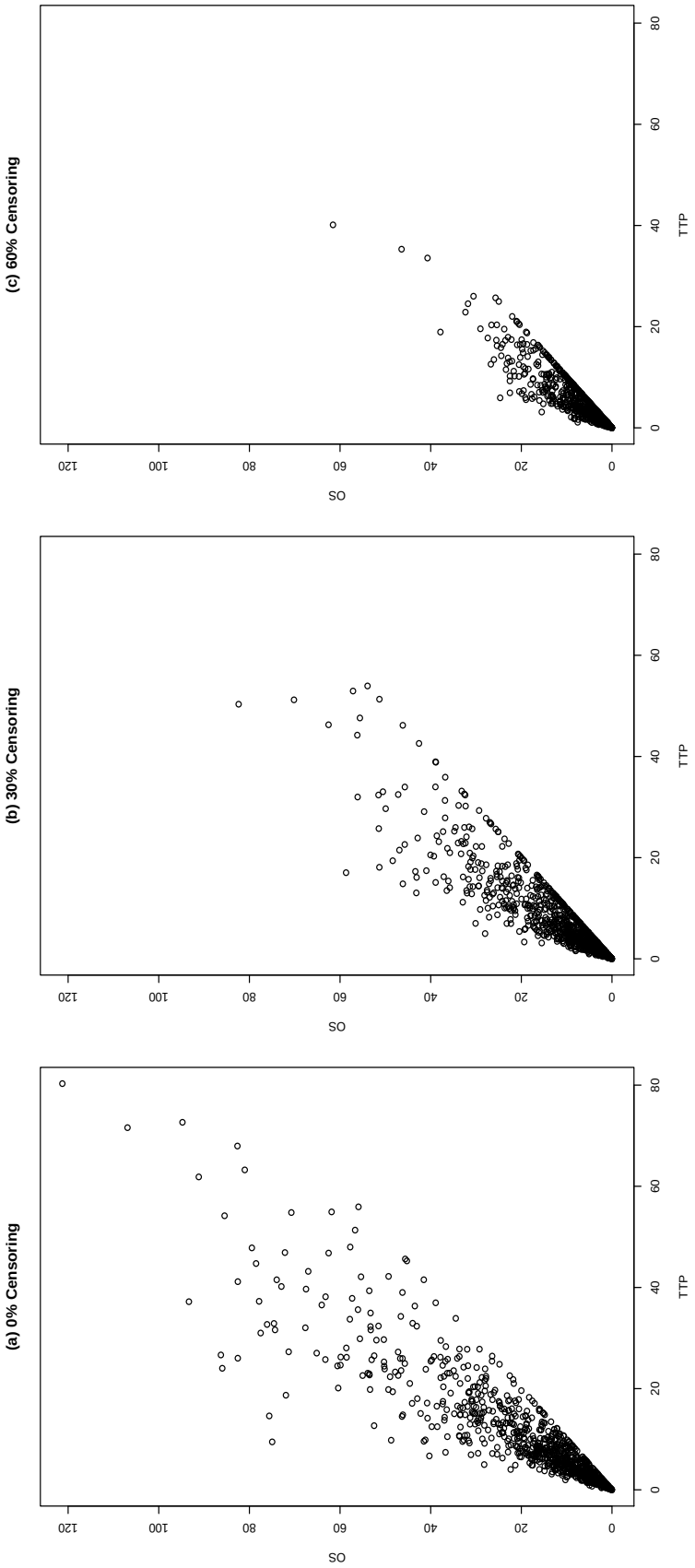

Figure 3.11: Scatterplot of 1,000 values of S (TTP) and T (OS) generated from the Gumbel copula ($\tau = 0.8$)

3.4.3 Endpoint Symmetry

Arguably, the most impactful finding of the current study comes from the switch of surrogate endpoint from TTP to PFS. Whilst the highest underlying value of τ appears mostly unaffected by this, the lowest value (0.2) was substantially over-estimated across all scenarios when based on PFS. Results of this simulation study have demonstrated that truly low values of individual-level association between PFS and OS cannot be reliably identified, and this is considered to be a result of the composite nature of the PFS endpoint. For this study, datasets were simulated to provide a date of progression and date of death for each patient, with the required strength of association between these two separate outcomes. In re-defining a surrogate endpoint to be a combination of the two events, such that it also includes data from the true endpoint, it is reasonable to assume that the intended association between PFS and OS would be higher than that between TTP and OS for the same set of data values. Therefore, values of τ could be expected to be over-estimated by the two-stage meta-analytic copula method when based on such a composite endpoint. However, this is the approach that is taken to analyse PFS as an endpoint, and which would be reflected if PFS were used as a surrogate endpoint.

3.5 Implications of Results

The simulation study has revealed a number of topics that require consideration before implementing the two-stage meta-analytic copula method to evaluate a surrogate endpoint. In particular, it must be considered how the method can be practically applied in the setting of PFS, and how to determine whether such a surrogate endpoint is truly reliable for future use. Such practical implications are now described, followed by some recognised limitations of the simulation study conducted, and further work.

3.5.1 Practical Implications

The results of the simulation study described in this chapter have raised one very important issue that is not otherwise identified in the current literature; the low reliability of the method when applied to endpoints that are not symmetrical and/or that incorporate data from the true endpoint, such as PFS. The poor performance of the two-stage meta-analytic copula method in estimating individual-level surrogacy when the proposed surrogate endpoint incorporates data from the true clinical endpoint is evident, yet this key assumption of the copula model has previously been ignored when applying the approach to the assessment of PFS as a surrogate for OS. This clearly has the potential for severe consequences. None of the previous simulation studies conducted in the literature have addressed this question, yet the results here have demonstrated high cause for concern.

Secondly, it has been shown that the method cannot be considered uniformly applicable when there exist only a small amount of data on which to base a surrogacy evaluation. Whilst this is expected for the estimation of trial-level surrogacy, the large variability present in estimates of individual-level surrogacy (with the exception of symmetrical surrogate and true endpoints under correctly specified models) indicate that applications of the method to such small samples may lead to incorrect conclusions, and inappropriate use of a surrogate endpoint.

The third and final aim of the simulation study was to determine whether the methodology could reliably identify truly low levels of association between surrogate and true endpoints. The results demonstrated that this is possible for the case of symmetrical endpoints under correct model specification, albeit with higher variability than higher levels of association, but that reliability of estimates deteriorates as model assumptions are violated. Such violations can be in the form of non-symmetry of surrogate and true endpoints, or a change in the dependence structure of the data. As a result, use of the two-stage meta-analytic copula method in such scenarios is likely to lead to results that over-estimate the strength of association between endpoints, and incorrect conclusions that

a surrogate is reliable enough to use in future clinical trials. The importance of ensuring that the copula model being used to estimate the parameters adequately fits the data is therefore critical, consistent with the finding of Renfro et al. (2015). This can be achieved through inspection of Akaike's Information Criterion (AIC) (Akaike, 1974) or Schwarz's Bayesian Information Criterion (BIC) (Schwarz, 1978), where both criteria penalise the (log) likelihood by a value linked to the total sample size, n, of all clinical trials included in the meta-analysis, and the number of parameters in the model, p. The AIC is penalised by subtraction of p from the log-likelihood value, whereas the BIC penalises by subtracting a value of $\frac{p\log(n)}{2}$. A higher value for either criterion therefore indicates a better fit.

In addition to the above, the complex joint modelling required by the two-stage meta-analytic copula method leads to many issues of non-convergence when using PFS as a surrogate for OS. The lack of non-convergence from the TTP scenarios suggests that this is caused by the change in surrogate endpoint, and in some cases the values are so high that more than half of the simulation runs failed to provide estimated values of τ and R^2_{trial}. This implies that, in practice, it may be very difficult to apply the method, particularly when the true underlying individual association is low.

3.5.2 Limitations of the Simulation Study

Although investigating a large number of scenarios, the simulation study described in this chapter is subject to some limitations. Firstly, data generation was based on two different copula models, one that was consistent with the analysis approach and one that was not. The purpose of using two different copulas was to assess the impact on performance when the underlying data structure does not match that assumed by the model. However, since both approaches were based on a copula function, as is used by the surrogacy evaluation approach, this may have had the potential to bias the results in a favourable way. To address this concern, further examination of the two-stage meta-analytic copula method was conducted using data that was generated using a bivariate lognormal distribution,

and the results of this third data generation algorithm indicate that the approach of using Clayton and Gumbel data generation did not bias the findings. Results from the lognormal data generation can be found in Appendix A, Figures A.9 and A.10, and a brief description of the data generation procedure is provided in Section 4.2.2.

A second limitation is that all scenarios considered the same treatment effect in TTP/PFS (hazard ratio ≈ 0.67) and OS (hazard ratio ≈ 0.82) between experimental and control arms. These values were selected to reflect real-life scenarios, where the time required to run a new study could be substantially shorter if based on the surrogate endpoint rather than the true endpoint. Although these treatment effects were held constant across all simulations, Burzykowski (2001) considered variation in treatment effects from no difference (on S or T) to hazard ratios of 0.67 for both surrogate and true endpoints, and observed no major differences in results. Further, the strength of association defined by the copula data generation is fixed prior to selection of the treatment effects. Monotonic transformation of the event times to reflect the treatment effects therefore does not impact the underlying values of τ. Based on this structure and the findings of Burzykowski (2001), no change in conclusion is therefore expected if the treatment effects were to be varied.

A third limitation is that the simulation study conducted and described in this chapter considers use of only one copula model when evaluating the potential surrogate endpoint; the Clayton copula. As described by Burzykowski et al. (2001), any choice of copula could be used. This choice should be based on the best fit to the data, and so results may vary slightly when using another copula function. However, the use of two different copula models in the data generation process demonstrates how the method performs under correctly and incorrectly specified models, and it is considered unlikely that use of a different copula under these two scenarios would give results that dramatically change the findings.

A recognised issue with the two-stage meta-analytic copula method is the potential bias in the estimation of R^2_{trial} from the use of estimated treatment effects in Stage 1 of the

modelling approach without correction for estimation error. Whilst adjusted estimators have been proposed to correct for this issue, this simulation study used the unadjusted estimators only. The reason for this approach is that the investigation of Burzykowski (2001) demonstrated that the application of the adjusted estimators is hampered by issues with convergence. These issues were considered severe enough to conclude that the adjusted measures cannot be used in practice (Burzykowski et al., 2005). Hence, the unadjusted estimators are considered the most appropriate method with which to assess the performance of the two-stage meta-analytic copula method via simulations. Of note, while the unadjusted estimates cannot range outside of the unit interval $[0, 1]$, confidence intervals of these unadjusted estimates can take values outside of this range due to the approximation used in the delta method to calculate the parameter variance. When this occurs, it is recommended to truncate the confidence intervals to the admissible range. Adjusted estimators, based on the methods proposed by Burzykowski et al. (2005), can take values outside of the unit interval, further hampering their interpretation.

Finally, what this simulation study has shown is that having up to six clinical trials with data available for analysis is insufficient to provide a reliable assessment of trial-level surrogacy. Findings from Burzykowski (2001) suggest that having $10-20$ trials is sufficient for this purpose, and so the question remains as to the minimum number of trials (> 6 but ≤ 10) that could be considered necessary for an evaluation of this association. One important element to this question is whether the underlying individual and trial-level strengths of association are consistent across multiple trials, which may be dependent on the patient population, trial design and treatment under investigation.

3.5.3 Further Work

The findings from the simulation study presented in this chapter suggest that the complex joint modelling of surrogate and true endpoints can lead to difficulty in implementing the two-stage meta-analytic copula method. Further, when the method provides estimates of

trial and individual-level surrogacy, these can be misleading when the strong assumptions of the copula models being used are not appropriate, in particular under non-symmetry of surrogate and true endpoints. Whilst the latter has been demonstrated for the first time in this research, the former is a known concern, which many have tried to address through the use of unified approaches to evaluating surrogacy.

A number of unified approaches have been described in Section 2.4, one of which shows promise and has been recommended as the preferred choice of measures for the evaluation of surrogacy (Ensor et al., 2016). This information theory method, described in Section 2.4.2, is therefore investigated further in the next chapter, with the aim to determine whether this approach can reduce the complexity of the modelling process, and provide more reliable estimates of surrogacy for the PFS setting.

Chapter 4

A Unified Approach Based on Information Theory

4.1 Introduction

A desirable feature of statistical methodology designed to evaluate surrogate endpoints is that there is consistency in interpretation of the results across different endpoint types. In particular, when researchers are investigating multiple potential surrogates with a desire to determine which may be the most reliable for future use, it is important that the surrogacy measures, be they of an R^2 type or otherwise, are reflecting the same underlying concepts. This avoids a situation where different conclusions may be drawn purely due to the choice of statistical methodology. As was described in Section 2.3, many of the methods currently available to assess potential surrogates have different assumptions and modelling structures depending on the endpoints under investigation, 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.

In a recent systematic review of surrogate endpoint methodology, Ensor et al. (2016) recognise this and note that the information theory method described in Section 2.4.1 has an advantage over the two-stage meta-analytic approach in that it offers a unified

interpretation. The time-to-event application of the approach (Section 2.4.2), offers a number of further potential advantages. Firstly, there is no need to define complicated joint distributions; measures are based on parameters of individual models which are available in many standard software packages. Secondly, there are no assumptions of endpoint symmetry, something that has been shown to adversely impact surrogacy assessment when based on a copula modelling approach. The same underlying theory is also applicable to both individual-level and trial-level surrogacy, as well as being applicable across all types of surrogate and true endpoints. Finally, it has been suggested that the information theory method can also be considered appropriate for the evaluation of time-ordered endpoints such as exploration of PFS as a surrogate for OS (Pryseley et al., 2011).

Pryseley et al. (2011) conducted a simulation study to assess a number of different approaches to estimate surrogacy parameters based on the information theory concept, however there are a number of relevant topics that remain unexplored. As a result, this chapter contains details of a simulation study designed to examine these previously unexplored areas. Firstly, it is of interest to determine whether the performance of the information theory method is impacted when measures are based on a meta-analysis of similar clinical trials, rather than being based on one trial only. This also allows estimation of trial-level association, and to be consistent with the two-stage meta-analytic copula investigation in Chapter 3, a linear relationship between treatment effects is assumed. Secondly, since there is no need for definition of the joint distribution between endpoints, it is of interest to assess whether the method is sensitive to the underlying data structure. Further, given the over-estimation of the two-stage meta-analytic copula method when evaluating the commonly used endpoint of PFS, it is of interest to determine whether the information theory approach would be able to provide more reliable results for the surrogacy of this endpoint against a true endpoint of overall survival. The robustness of the results from the simulation study is also improved as compared to that of Pryseley et al. (2011) through the use of a tenfold increase in the number of simulation runs ($5,000$ compared to 500).

4.2 Simulation Study

4.2.1 Choice of Data Generation Procedure

As noted previously, 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 association, R_h^2, make this very difficult, as they are estimated from conditional models using the likelihood ratio (see Equation (2.4)), meaning that each sample would have a slightly different underlying R_h^2. The measure of surrogacy is estimated using (conditional) model coefficients for both the surrogate and treatment (from two models), as well as the Kaplan-Meier survival estimates for each sample, and it would be very difficult to simultaneously control each of these parameters such that the overall strength of association was preserved.

Previous investigations of the R_{XOQ}^2 measure (described in Section 2.4.2) were based on datasets simulated according to a Cox proportional hazards model (Xu and O'Quigley, 1999). In this set-up, one covariate was included in the model and the coefficient of this single covariate, β, was used to control the overall strength of association between the covariate and outcome. Whilst this allows for specification of the strength of this relationship, it is not clear what value of covariate coefficient would be reflective of 'poor', 'medium' or 'strong' surrogacy, since there is no bound on the range of values that can be selected. Further, the strength of covariate coefficient may depend on the disease setting, or be impacted by other covariates in the model, such as treatment. There could therefore be considerable subjectivity in the selection of coefficient values, and the impact of this on the resulting estimation of R_{XOQ}^2 is not currently clear.

Instead, the study of Pryseley et al. (2011) was based on data generated using a Clayton copula function, controlling association between endpoints through the copula parameter,

τ. Whilst the value of τ may not perfectly reflect the true information theoretic measure of association, such an approach allows for overall control of the true individual (τ) and trial (R^2_{trial}) association levels, subject to sample variability, and conveniently allows indirect comparison with results of the two-stage meta-analytic copula method. The strength of surrogacy is also unaffected by other covariates in the model, as these are incorporated later and do not affect the underlying value of τ.

Despite the potential limitations of conducting a simulation study of the information theory approach based on data generated using a copula model, it is considered important to understand how the estimates of individual and trial-level surrogacy would compare between the two-stage meta-analytic and information theory methods when applied to the same datasets. It would seem relevant that any practical application of surrogacy evaluation would be based on a number of different methods, including some sensitivity analyses. A substantial limitation to interpretation of surrogacy would occur if the available methods gave conflicting results for the same dataset. As such, the information theory method was applied to the identical datasets generated as part of the simulation study of the two-stage meta-analytic copula method described in Section 3.2, for the same scenarios. As a reminder, these scenarios are displayed in Table 4.1.

Despite the information theory method not making assumptions around the joint dependency structure between S and T, both Clayton and Gumbel copula generated datasets were re-used such that an assessment of the sensitivity of the approach to the dependency could be made. To examine whether use of the copula-generated data could lead to bias in estimation of information theory surrogacy measures, selected scenarios were also run using data generation that did not involve a copula model, and this process is described further in the next section.

Table 4.1: Simulation Scenarios

Factor	Scenarios under simulation
Surrogate Endpoint	TTP, PFS
Data Generation	Clayton, Gumbel
Number of trials	4, 6
Number of patients per trial	80, 120, Mixed (50% each at $n = (80, 120)$)
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%
Range of treatment effects*, σ	0.1, 0.2

*Hazard ratios ranging $42\% - 203\%$ and $31\% - 238\%$ from the mean for $\sigma = 0.1, 0.2$ respectively.

4.2.2 Lognormal Data Generation

One alternative approach to data generation that does not employ a copula function is based on a bivariate lognormal distribution, previously used in the context of comparing estimators of Kendall's τ (Hsieh, 2010). Using this method, the parameter that controls the association between endpoints remains as Kendall's τ, which is transformed and used within the covariance matrix between S and T. Although this method also makes use of the joint distribution between the two endpoints, it does not use a copula model, and therefore provides a data generation algorithm that can adequately control the association parameter, but is not reliant on the choice of copula. In addition, the association between endpoints can be controlled using the same parameter as that used in copula data generation, τ, with the same values (0.2, 0.5, 0.8). This alternative approach was used to explore whether the results of the information theory method of estimating individual-level association are consistent with those based on copula-generated data. Since exploration of the lognormal data generation approach was considered a sensitivity analysis to confirm

whether any bias may have been introduced through use of the copula generation, only selected scenarios were re-run, shown in Table 4.2.

Table 4.2: Selected Simulation Scenarios for Lognormal Data Generation

Factor	Scenarios under simulation
Surrogate Endpoint	TTP, PFS
Number of trials	6
Number of patients per trial	120
Trial-level association	0.5
Individual-level association	0.2, 0.5, 0.8
Censoring Rate (on T)	0%, 30%
Range of treatment effects*, σ	0.1

*Hazard ratios ranging $42\% - 203\%$ from the mean.

For lognormal data generation, surrogate and true endpoint values S_{ij} and T_{ij} need to be generated for each patient j from trial i, with required underlying individual and trial level association. This is consistent with the copula data generation as described in Sections 3.2.4 and 3.2.5. To generate these values, a three-step process was created, which was able to also incorporate the range of simulation parameters listed in Table 4.1 and allow consistency with previous simulated datasets.

Step 1:

First, two variables, $\log(S_{ij}^0)$ and $\log(T_{ij}^0)$, were generated from a bivariate Normal distribution with

$$\begin{pmatrix} \log(S_{ij}^0) \\ \log(T_{ij}^0) \end{pmatrix} \sim N_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right],$$

where ρ is Spearman's correlation between $\log(S_{ij}^0)$ and $\log(T_{ij}^0)$, reflecting Kendall's τ through the relationship between these two parameters; $\rho = \sin\left(\frac{\tau\pi}{2}\right)$ (Kruskal, 1958). As

such, values of τ of 0.2, 0.5 and 0.8 are represented by correlation values of 0.309, 0.707 and 0.951, respectively.

Step 2:

In order to generate time-to-event data with the required characteristics, the second step of the process is to rescale the values of $\log(S_{ij}^0)$ and $\log(T_{ij}^0)$ to ensure that the final distributions of the generated S_{ij} and T_{ij} are similar to those generated using the copula models. This is achieved by transforming using a specified mean and variance, which can be considered equivalent to applying selected baseline hazard functions when converting copula-generated uniform endpoints to have exponential distributions. Since the parameters $\log(S_{ij}^0)$ and $\log(T_{ij}^0)$ follow a bivariate Normal distribution, the transformation based on the mean and variance follows the standard process for converting from a standard Normal distribution with mean of 0 and variance of 1 to a Normal distribution with chosen mean and variance, using

$$\begin{aligned} S_{ij}^* &= m_S + \left(\sqrt{\sigma_S^2}\log(S_{ij}^0)\right), \\ T_{ij}^* &= m_T + \left(\sqrt{\sigma_T^2}\log(T_{ij}^0)\right), \end{aligned}$$

where m_S, m_T are the means and σ_S^2, σ_T^2 the variances of the required distributions of S_{ij}^* and T_{ij}^* respectively. The choice of these mean and variance values is described below, after Step 3 of the data generation algorithm.

Step 3:

The third step of the process is to transform the values S_{ij}^* and T_{ij}^* by exponentiating, to obtain lognormally distributed time-to-event values. In addition, the effect of treatment and the underlying trial-level association value need to be incorporated. Since a restricted selection of simulation parameters are being explored for this data generation method, trial-level association is held fixed at a value of 0.5 to demonstrate a 'medium' level of association. In order to incorporate the required treatment effect, as well as take into consideration the underlying trial-level association, the same approach as used for copula simulation was applied. This was achieved by multiplying the lognormally distributed

time-to-event variables according to

$$S_{ij} = \exp(\mu_{S_i} + (\alpha + a_i)Z_{ij}) \exp(S^*_{ij}), \quad (4.1)$$

$$T_{ij} = \exp(\mu_{T_i} + (\beta + b_i)Z_{ij}) \exp(T^*_{ij}), \quad (4.2)$$

where Z_{ij} denotes treatment and takes a value of 0 or 1, $\alpha = -0.4$ and $\beta = -0.2$ represent the treatment effects (HR of 0.67 and 0.82 for S and T respectively), and μ_{S_i}, μ_{T_i}, a_i and b_i are random effects introduced to control the underlying trial-level association. Further details of the derivation of these random effects are described in Chapter 3. This approach of multiplying the generated outcome time by the exponential terms including treatment and random effects is identical to the approach used in the copula-generated data to control these parameters.

In order to ensure that the datasets simulated according to this algorithm were comparable to those generated using the Clayton and Gumbel copula functions in terms of the summary statistics, the mean and variance parameters, m_S, m_T, σ^2_S and σ^2_T were selected based on the observed values of these parameters in the copula-generated datasets. To estimate these, the random variables S_{ij} and T_{ij} derived from the copula model in Equations (3.2) and (3.3) were equated to those in Equations (4.1) and (4.2) above, leaving

$$\begin{aligned} -\lambda_S^{-1} \exp(\mu_{Si} + (\alpha + a_i)Z_{ij}) \log(S^{0c}_{ij}) &= \exp(\mu_{S_i} + (\alpha + a_i)Z_{ij}) \exp(S^*_{ij}), \\ -\lambda_T^{-1} \exp(\mu_{Ti} + (\beta + b_i)Z_{ij}) \log(T^{0c}_{ij}) &= \exp(\mu_{T_i} + (\beta + b_i)Z_{ij}) \exp(T^*_{ij}), \end{aligned}$$

where $\log(S^{0c}_{ij})$ and $\log(T^{0c}_{ij})$ denote the values of $\log(S^0_{ij})$ and $\log(T^0_{ij})$ based on the copula model in Equations (3.2) and (3.3), λ_S and λ_T are the baseline hazards selected for the copula data generation, and the vector of random effects, $\exp(\mu_{Si} + (\alpha + a_i)Z_{ij})$, contains the trial-specific parameters as described previously. Since this vector of random effects in both sides of the equations is the same, it follows that

$$\begin{aligned} \log(-\lambda_S^{-1} \log(S^{0c}_{ij})) &= S^*_{ij}, \\ \log(-\lambda_T^{-1} \log(T^{0c}_{ij})) &= T^*_{ij}, \end{aligned}$$

and so the mean and variance of the left hand side of these equations can be used to provide an estimate of the mean and variance of S^*_{ij} and T^*_{ij}. Values of the mean and variance were therefore considered from both the Clayton and Gumbel generated values of $\log(S^{0c}_{ij})$ and $\log(T^{0c}_{ij})$, and were selected as $m_S = 1.1$, $m_T = 2.1$, $\sigma^2_S = \sigma^2_T = 1.6$. Since Kendall's τ is based on ranks of variables rather than specific values, the simulated value of τ is unaffected by such a monotonic transformation.

4.2.3 Modelling Structure

The information theory method of evaluating surrogacy is based on individual models of the true outcome, one with treatment only and one with treatment and surrogate as covariates. The estimated coefficients of these covariates are then used to compare the two models to determine how well the addition of the surrogate in the model can improve the model predictions. In order to include the surrogate endpoint in the model, it is necessary to express it as a time-dependent covariate, to reflect that it is measured post-baseline and to appropriately estimate the respective model coefficient. This is achieved by splitting the time period from baseline to the true endpoint, $[0, T)$, into two intervals that reflect the potential change in surrogate outcome at time S; $[0, S)$ and $[S, T)$. During the first interval there is no progression, and during the second interval there may or may not be disease progression dependent on the disease status at time S. Time-dependent indicator variables are used to denote disease status both at time T and for the surrogate outcome.

When TTP is used as the surrogate, the covariate status during the interval $[S, T)$ is based only upon the disease status provided by the surrogate endpoint at time S. Therefore, if a patient experiences disease progression at time S, this is accounted for in parameter estimation through a change in the value of the time-dependent covariate from zero to one. If a patient does not experience disease progression during their period of observation, their time-dependent covariate remains at a value of zero across the entire interval $[0, T)$.

For PFS, the set-up is slightly different due to the need for PFS to reflect both disease progression and death. In particular, when a patient experiences death without prior disease progression, the value of S would be truncated to the value of T, and both endpoints would be considered as events. In a time-dependent covariate setting, this equality of time values would lead to the interval $[S, T)$ having zero length, and so the surrogate outcome not being accounted for. To avoid this, in cases where patients had death without prior progression, the interval $[S, T)$ was assumed to have length of one day, such that the data reflects the surrogate outcome.

An alternative representation of PFS as the surrogate endpoint is also suggested by Pryseley et al. (2011), although not explored in their study. Given the non-symmetric nature of endpoints used as potential surrogates for survival, the suggestion is that the outcome T could be replaced by post-progression survival, $T - S$. The restriction that S must be shorter than T can then be immediately implemented within the information theory approach. This offers a benefit over the two-stage meta-analytic copula method, where the endpoint symmetry cannot be easily adjusted for. As a result, further investigation of this approach was also considered, where the outcome T was replaced with a value of $T - S$, and the surrogate endpoint S was considered as a covariate that was no longer required to be time-dependent. This is the first known evaluation of this key advantage of the information theory approach.

To estimate trial-level association, treatment effects on S and T were estimated through the use of separate proportional hazards models; one with T as outcome and treatment as the only covariate, and one with S as the outcome and treatment as the only covariate. A linear relationship between these estimated treatment effects on S and T is assumed, and the square of the correlation coefficient is used as an estimate of R^2_{trial}. Using a normal linear model for association between treatment effects on S and T, this estimated trial-level association has an information theoretic interpretation, as described by Alonso and Molenberghs (2007).

4.3 Results

Surrogacy measures based on the information theory approach will be denoted $R^2_{h,i}$ and $R^2_{h,t}$ for individual-level and trial-level surrogacy respectively. Results are presented for each of the factors described in Table 4.1, first for TTP (Section 4.3.1) and then for PFS (Section 4.3.2). Similar to the two-stage meta-analytic copula method, results between values of the parameter controlling the ranges of treatment effects across trials ($\sigma = 0.1, 0.2$) were comparable, hence only the results for the smaller ranges are presented herein; remaining results can be found in Appendix B (Figures B.2 to B.9) for both individual and trial-level surrogacy.

Presentation of results based on the Clayton and Gumbel copula-generated data is consistent with that of the previous chapter. Each scenario is displayed in a figure containing nine individual plots, showing all combinations of the investigated values of individual and trial-level surrogacy; each row contains a fixed individual-level value, and each column a fixed trial-level value. Within each of these nine individual plots are all results for the fixed combination of τ and R^2_{trial} across all numbers of trials (N=4, 6), patients within each trial (n=80, 120, mixed) and proportions of censoring (0%, 30%, 60%). Within the plots, estimates considered to be outliers are not presented (values are considered outliers if they lie below the first quartile or above the third quartile by a margin of 1.5 times the inter-quartile range). To support the graphical displays, summary tables are included to show the median percentage bias across all simulation runs (calculated as the percentage difference between the estimated value of $R^2_{h,i}$ or $R^2_{h,t}$ and the respective value used in data generation, τ or R^2_{trial}, as a proportion of the value used in data generation). To improve readability, only the largest sample sizes are included in these summary tables ($N = 6$, $n = 120$). Results of the scenarios selected for further investigation using lognormal generated data are presented separately in Figures 4.3 (TTP) and 4.8 (PFS). These figures include three individual plots of the estimates of $R^2_{h,i}$ for varied underlying τ and a fixed underlying value of $R^2_{trial} = 0.5$. In these additional figures, results from the Clayton and

Gumbel copula-generated data are also included for easier comparability.

Overall, there were very few issues with model convergence for the information theory method, with non-convergence of 0.04% for TTP and 0.14% for PFS. This reflects the simpler computational properties of the method as compared to the two-stage meta-analytic copula method, and demonstrates that the method is much easier to implement in practice.

4.3.1 Time-to-Progression

Estimation of $R^2_{h,i}$

Estimated values of $R^2_{h,i}$ for the Clayton generated data based on TTP are presented in Figure 4.1. Horizontal dashed lines represent the true value of τ used in data generation, and whilst the true value of $R^2_{h,i}$ may not perfectly match τ, it is important to understand whether the approach can reliably identify 'poor' from 'good' surrogates, and provide estimates that are broadly comparable to the underlying association within the data. Confidence intervals around the estimates of $R^2_{h,i}$ for the example scenario of TTP with Clayton generated data ($R^2_{trial} = 0.5$, $N = 4$, $n = 80$) are presented in Appendix Figure B.1 for information, but are not discussed further herein.

It is clear from the results of the TTP Clayton data that the information theory method consistently provides estimates of $R^2_{h,i}$ that are lower than the value of τ used in data generation, and this is true across all levels of individual and trial 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 magnitude of association. However, the main concern is the very large ranges of results, which increase as the true level of association increases. Under the lowest strength of association, estimates appear to have low variability, suggesting that truly poor surrogates could be reliably identified, however the estimates of medium ($\tau = 0.5$) and high ($\tau = 0.8$) strengths of association are widely spread, which hampers interpretation. Coupled with the overall

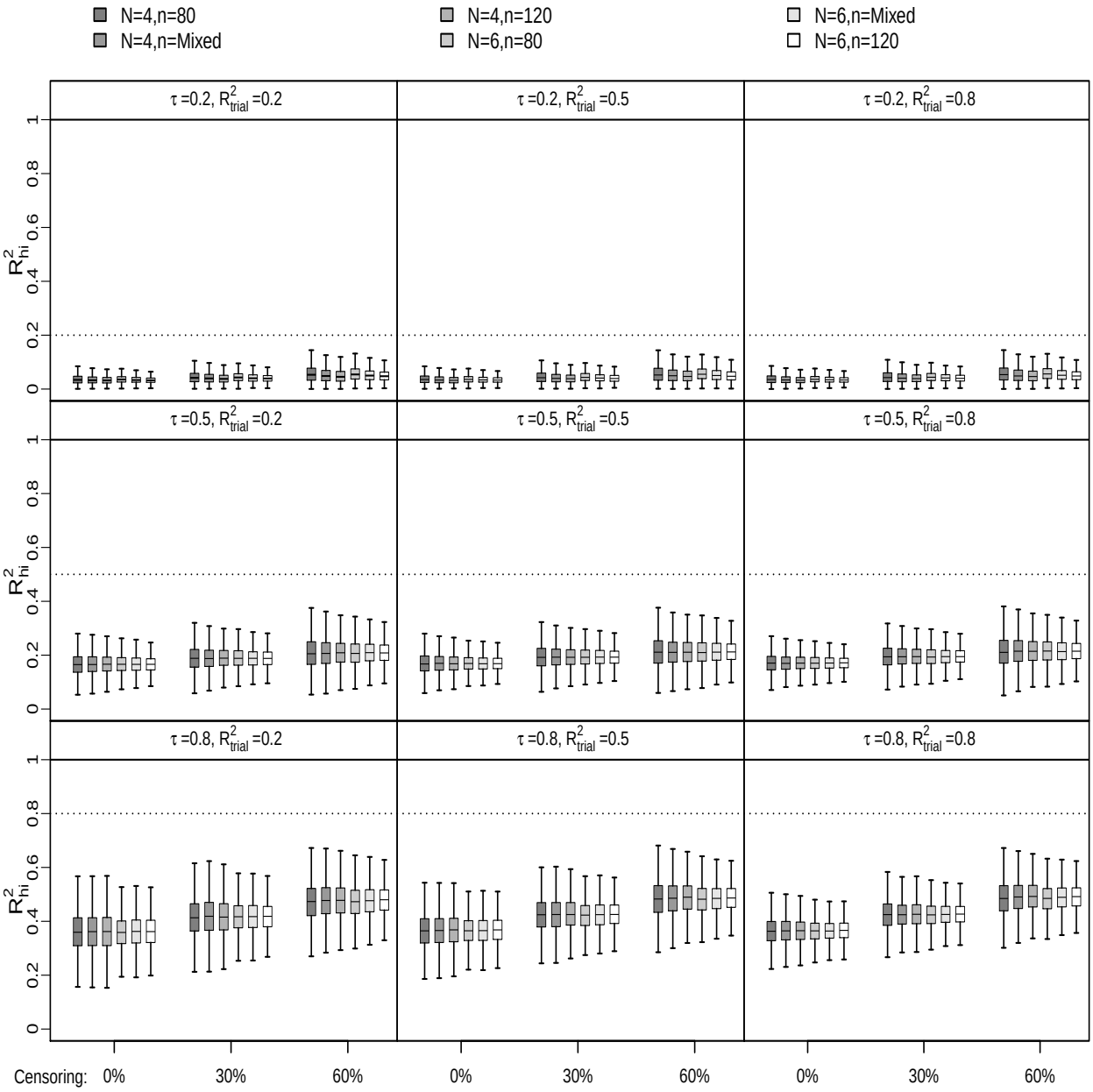

Figure 4.1: Boxplots of estimates of $R^2_{h,i}$: TTP, Clayton Copula Data Generation, Information Theory Application

lower estimates of individual-level association, it is unlikely that even the strongest surrogates could be identified, with estimates of $R^2_{h,i}$ 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, with similar results across most scenarios and only slight increases in estimates under the setting with highest τ. Further, as could be expected, the variability appears to improve slightly with increased number of trials and patients within trials. However, even under the largest sample sizes ($N = 6$, $n = 120$), the median bias remains large (Table 4.3), reaching as low as -84%. It is unlikely that individual pharmaceutical companies would have more available data than this, and so even with some allowance for $R^2_{h,i}$ to deviate from the true τ, it is clear that truly high surrogacy cannot be identified.

Since the information theory method does not employ a copula function in the modelling, it is not expected that results will vary substantially when switching to Gumbel data generation. However, it is of interest to understand whether the approach is sensitive to the dependence structure of the underlying data. Results of the method applied to Gumbel generated datasets are presented in Figure 4.2.

Overall, estimates are broadly consistent with those based on Clayton generated data. Across all scenarios, estimates of $R^2_{h,i}$ are lower than the value of τ used for data generation, but increase as the underlying strength of association increases, which is encouraging. However, as for the Clayton data, the large variability in results limits interpretation and does not allow for clear conclusions to be made, particularly when trying to identify strong surrogates (where estimates of $R^2_{h,i}$ range 0.2 to 0.8, median bias as low as -47%).

In addition, there are a few subtle changes in the results based on the Gumbel data. Values are very slightly higher than those based on the Clayton generated data, with median bias reducing by values up to 10% when no censoring is present. Whilst such marginal changes reflect stronger surrogacy, the increase is not sufficient to reflect the strength of association present in the data. More impactful, however, is the proportion of censoring within the datasets, with higher estimates of $R^2_{h,i}$ resulting from datasets

Table 4.3: % Bias of Estimates of $R^2_{h,i}$ and $R^2_{h,t}$: $N = 6$, $n = 120$, TTP with Clayton Data

τ	R^2_{trial}	% Censoring	Median % bias $R^2_{h,i}$	Median % bias $R^2_{h,t}$
0.2	0.2	0%	-83.744	-0.651
0.2	0.5	0%	-83.375	-30.177
0.2	0.8	0%	-83.192	-34.706
0.2	0.2	30%	-80.372	-6.649
0.2	0.5	30%	-80.425	-38.448
0.2	0.8	30%	-80.098	-43.224
0.2	0.2	60%	-76.185	-20.801
0.2	0.5	60%	-76.269	-54.325
0.2	0.8	60%	-75.758	-59.756
0.5	0.2	0%	-66.774	7.275
0.5	0.5	0%	-66.285	-15.243
0.5	0.8	0%	-65.683	-16.448
0.5	0.2	30%	-62.395	3.029
0.5	0.5	30%	-61.452	-21.479
0.5	0.8	30%	-61.083	-25.610
0.5	0.2	60%	-58.326	-4.899
0.5	0.5	60%	-57.590	-38.881
0.5	0.8	60%	-57.035	-46.030
0.8	0.2	0%	-54.776	18.594
0.8	0.5	0%	-53.984	-0.539
0.8	0.8	0%	-54.206	-3.189
0.8	0.2	30%	-47.687	14.707
0.8	0.5	30%	-46.827	-5.222
0.8	0.8	30%	-46.659	-10.005
0.8	0.2	60%	-40.003	11.091
0.8	0.5	60%	-39.188	-19.980
0.8	0.8	60%	-38.536	-25.438

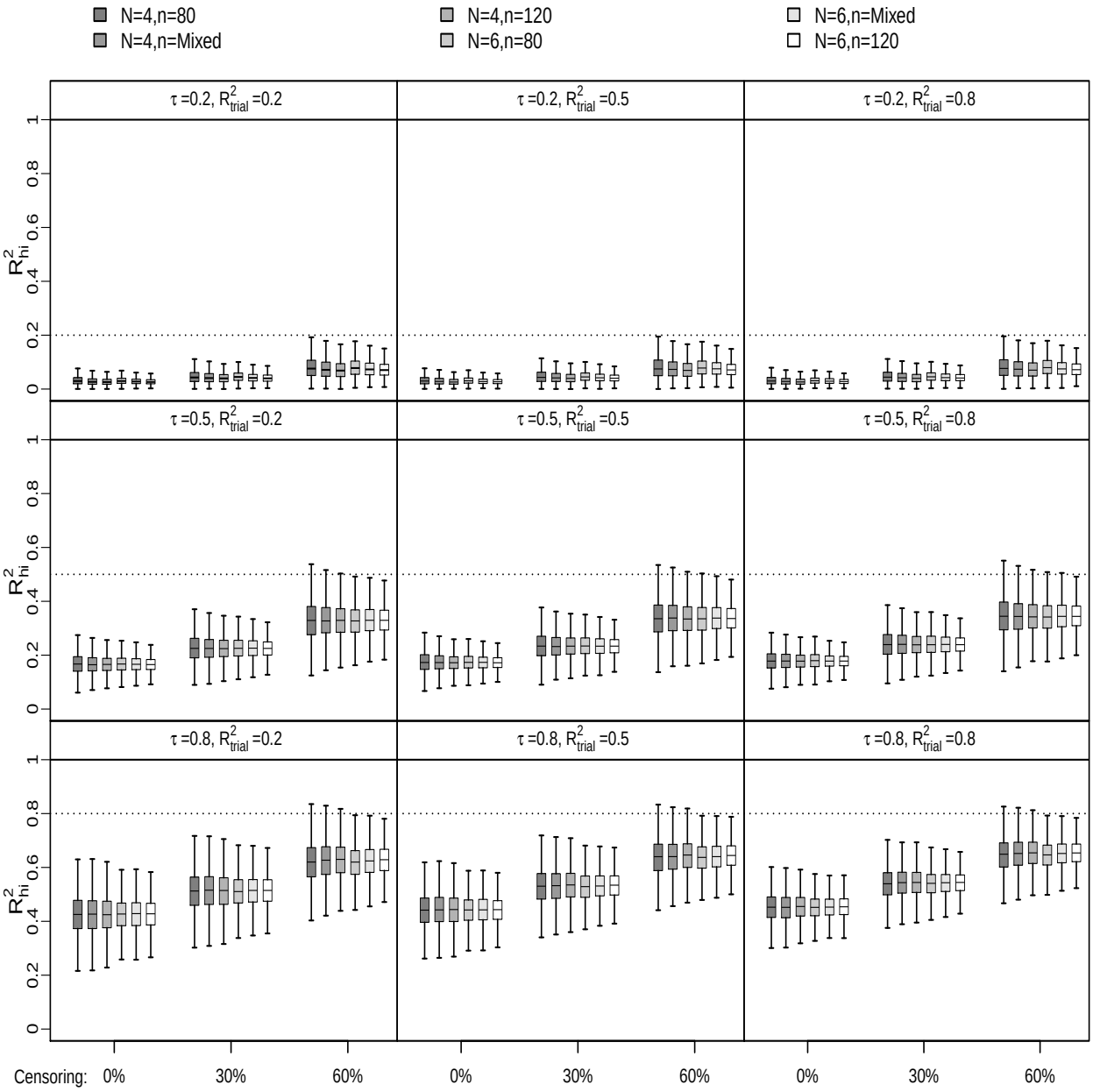

Figure 4.2: Boxplots of estimates of $R^2_{h,i}$: TTP, Gumbel Copula Data Generation, Information Theory Application

with the highest level of censoring (60%). Whilst there was no impact from censoring under Clayton data generation, these results suggest that the information theory method may be impacted in some circumstances, particularly for medium to high association levels. In addition, the increased censoring also appears to reflect larger variability for the low and medium association levels. Whilst increased sample size reduces variability, the improvement is not sufficient to allow reliable conclusions to be drawn when based on the small numbers of trials and patients investigated in this study.

Lognormal Data

As described previously, further simulations were conducted to assess the performance of $R^2_{h,i}$ in estimating individual-level surrogacy when a copula model was not used to generate the data. The lognormal data described in Section 4.2.2 was used to determine whether there was any bias introduced through use of the copula models, and was run $5,000$ times for a selected number of scenarios from the main simulation study. These scenarios considered both TTP and PFS, with $N = 6$ trials each containing $n = 120$ patients, trial-level association fixed at 0.5, with no censoring and under censoring of approximately 30%, for the smallest range of treatment effects. To explore the comparability of the final values of S_{ij} and T_{ij} to those generated from the copula models, histograms and summary statistics were used, which demonstrated high consistency between all methods.

Results for the TTP scenarios are presented in Figure 4.3. To aid interpretation, the results from the Clayton (light blue) and Gumbel (dark blue) copula functions for the same scenario are also included in the plots. Horizontal dashed lines represent the true value of τ used in data generation.

Results demonstrate some variability across the different data generation mechanisms, with estimates of $R^2_{h,i}$ from the Gumbel data being slightly higher than those from the Clayton data as previously noted. The white boxes, representing the estimates based on the lognormal data, appear to be consistent with those from the copula generated datasets,

Table 4.4: % Bias of Estimates of $R^2_{h,i}$ and $R^2_{h,t}$: $N = 6$, $n = 120$, TTP with Gumbel Data

τ	R^2_{trial}	% Censoring	Median % bias $R^2_{h,i}$	$R^2_{h,t}$
0.2	0.2	0%	-86.787	-9.456
0.2	0.5	0%	-86.455	-34.330
0.2	0.8	0%	-86.129	-40.796
0.2	0.2	30%	-80.134	-9.933
0.2	0.5	30%	-79.820	-40.837
0.2	0.8	30%	-79.456	-44.411
0.2	0.2	60%	-64.779	-10.309
0.2	0.5	60%	-64.479	-51.774
0.2	0.8	60%	-64.065	-56.653
0.5	0.2	0%	-66.971	-1.453
0.5	0.5	0%	-65.630	-25.284
0.5	0.8	0%	-64.377	-25.537
0.5	0.2	30%	-54.965	-2.139
0.5	0.5	30%	-53.276	-27.844
0.5	0.8	30%	-52.203	-30.784
0.5	0.2	60%	-34.101	-4.992
0.5	0.5	60%	-32.787	-36.197
0.5	0.8	60%	-31.196	-40.726
0.8	0.2	0%	-46.510	12.718
0.8	0.5	0%	-44.586	-4.658
0.8	0.8	0%	-43.269	-6.064
0.8	0.2	30%	-35.663	15.084
0.8	0.5	30%	-33.231	-7.275
0.8	0.8	30%	-31.960	-9.366
0.8	0.2	60%	-21.409	20.813
0.8	0.5	60%	-19.443	-11.171
0.8	0.8	60%	-18.286	-18.530

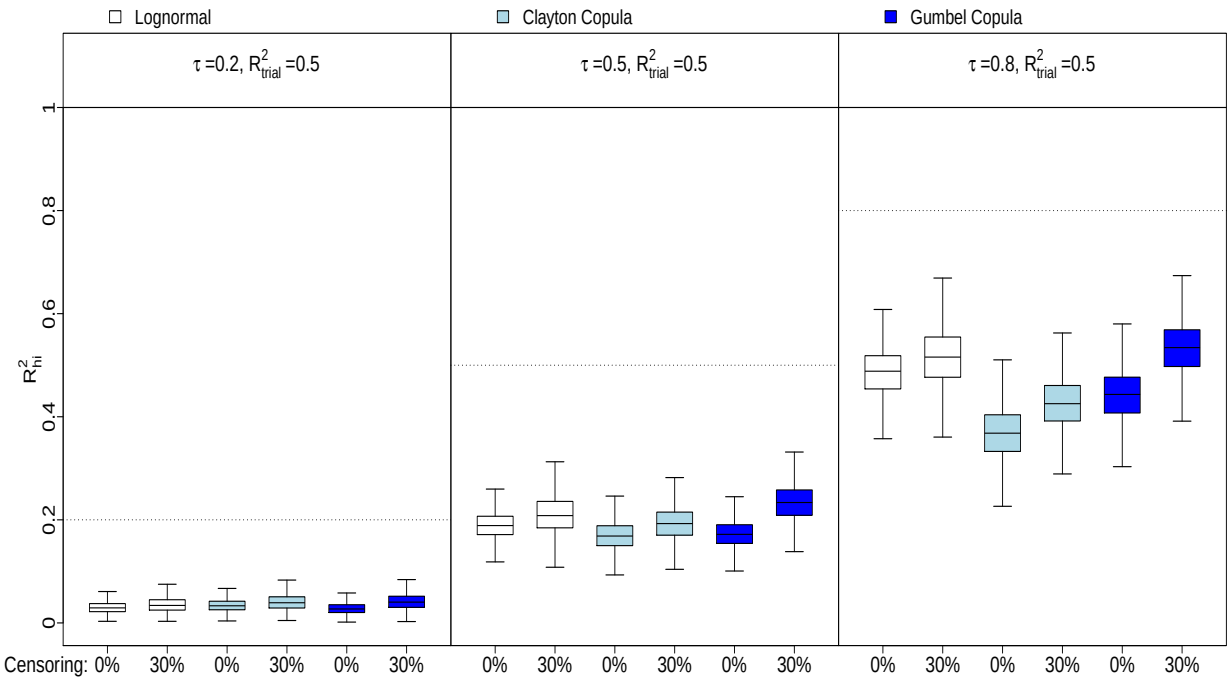

Figure 4.3: Boxplots of estimates of $R^2_{h,i}$: TTP, Information Theory Application to All Data Generation Methods (N=6, n=120)

being slightly higher than the Clayton-based estimates but similar to the Gumbel-based estimates. Variability in results appears also comparable to the copula-based estimates. Overall, these additional simulations confirm that there was no detrimental effect on the conclusions of the simulation study through the use of copula-generated data. Results have demonstrated that even general conclusions around the predictive strength of TTP as a surrogate for OS are difficult to make.

Estimation of $R^2_{h,t}$

Results of the estimation of trial-level association are presented in Figures 4.4 for Clayton-generated data and 4.5 for Gumbel generated data. Due to the similarity of results between the two data generation mechanisms, they will not be described separately here.

Overall, and as per expectations, estimates of $R^2_{h,t}$ were substantially variable, with values extending across the unit interval in almost all settings. Whilst there is a slight

upwards trend as the true R^2_{trial} increases, this increase is minor in comparison to the large variability, and does not allow for reliable conclusions. The increase in numbers of trials from 4 to 6 appears to reduce the variability, but not sufficient to consider this a large enough number of trials on which to base an assessment of surrogacy.

As was observed in results of the two-stage meta-analytic copula method (Section 3.3.1), there appears to be a link between the true value of τ and $R^2_{h,t}$, with higher estimates being observed when τ increases. As discussed previously, this is considered a result of the use of estimated treatment effects on S and T to calculate $R^2_{h,t}$, without any correction for estimation error. Whilst the procedure for estimation of treatment effects differs between the two-stage meta-analytic copula method and the information theory method, both use their respective estimates in calculation of trial-level association (R^2_{trial} or $R^2_{h,t}$). The issue caused by the potential under- or over-estimation of true correlation depending on the measurement errors therefore remains. Overall, results have demonstrated that when there are very few trials available, estimation of R^2_{trial} using the information theory method is poor.

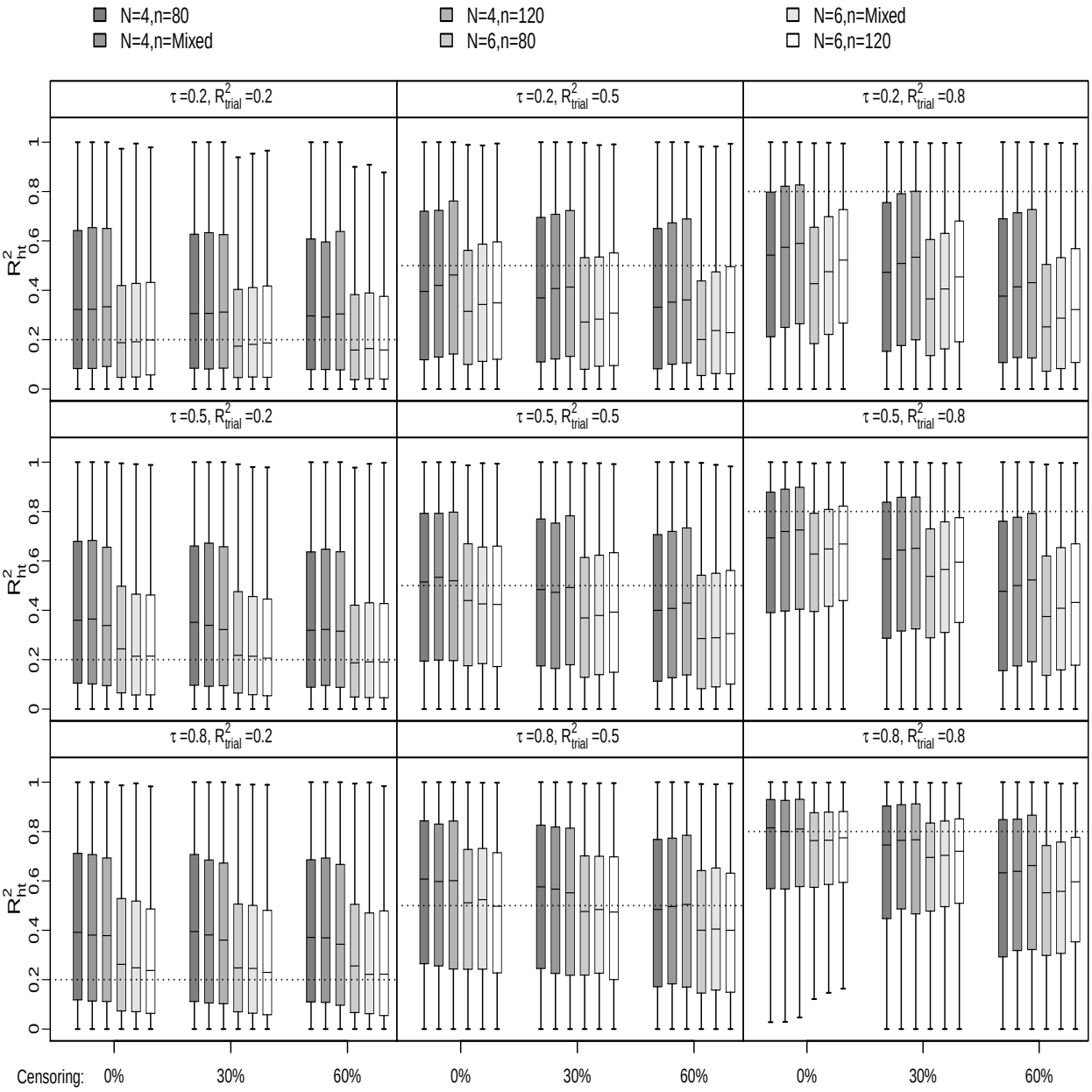

Figure 4.4: Boxplots of estimates of $R^2_{h,t}$: TTP, Clayton Copula Data Generation, Information Theory Application

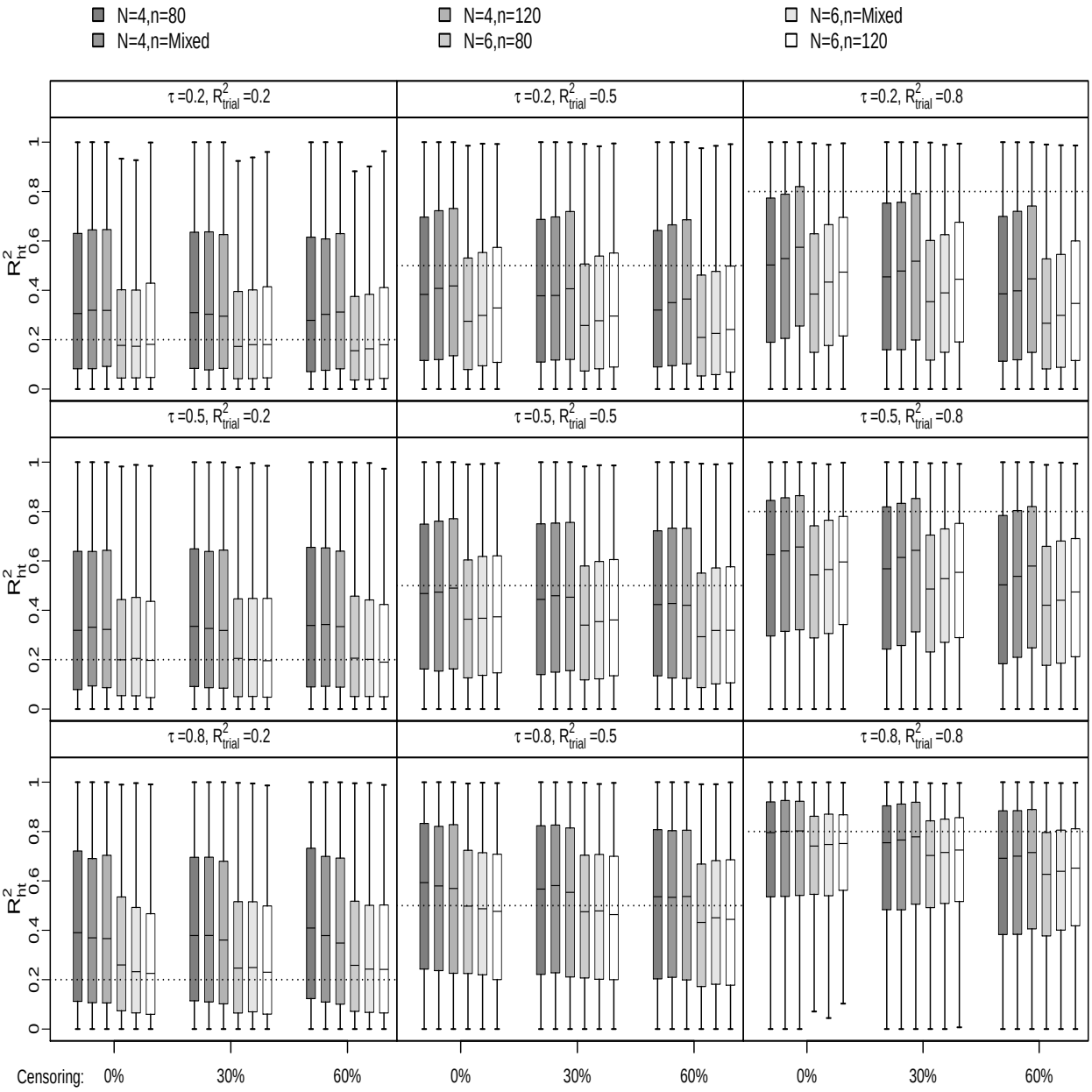

Figure 4.5: Boxplots of estimates of $R^2_{h,t}$: TTP, Gumbel Copula Data Generation, Information Theory Application

4.3.2 Progression-Free Survival

Estimation of $R^2_{h,i}$

Estimates of $R^2_{h,i}$ based on Clayton copula-generated PFS data are presented in Figure 4.6, with median bias for the largest sample sizes ($N = 6$, $n = 120$) provided in Table 4.5. Given that the information theory method does not assume any symmetry between the surrogate and true endpoint, it is anticipated that results would not be impacted by the change in surrogate endpoint to the same degree as those based on the two-stage meta-analytic copula method. In addition, the data structure used in the information theory approach is very similar between TTP and PFS, with the time-dependent covariate used to represent the surrogate differing only for patients who experience death without prior progression.

Results highlight strong similarities between TTP and PFS when based on Clayton generated data, with estimates of $R^2_{h,i}$ being reasonably comparable across the two endpoints. Values are marginally higher for the PFS data than the TTP data, likely due to those patients for whom the surrogate is impacted by the true endpoint (i.e. those who have death as the contributing PFS event). As for the TTP setting, the information theory method appears to be robust to the proportion of censoring in the data, except for the highest levels of association, where estimates increase slightly as the percentage of censoring increases.

Unfortunately, the large variability in estimates of $R^2_{h,i}$ is also present for PFS, particularly for medium to high levels of association. This means that for even the highest association, estimates of $R^2_{h,i}$ could be as low as 0.35, which could convince clinicians that the surrogate was not worthy of further consideration. The variability decreases slightly for larger sample sizes, but the bias remains at a value of up to -75% as compared to the value of τ used in data generation. The large negative bias together with the wide ranges of estimates that overlap between different underlying strengths of association hamper the

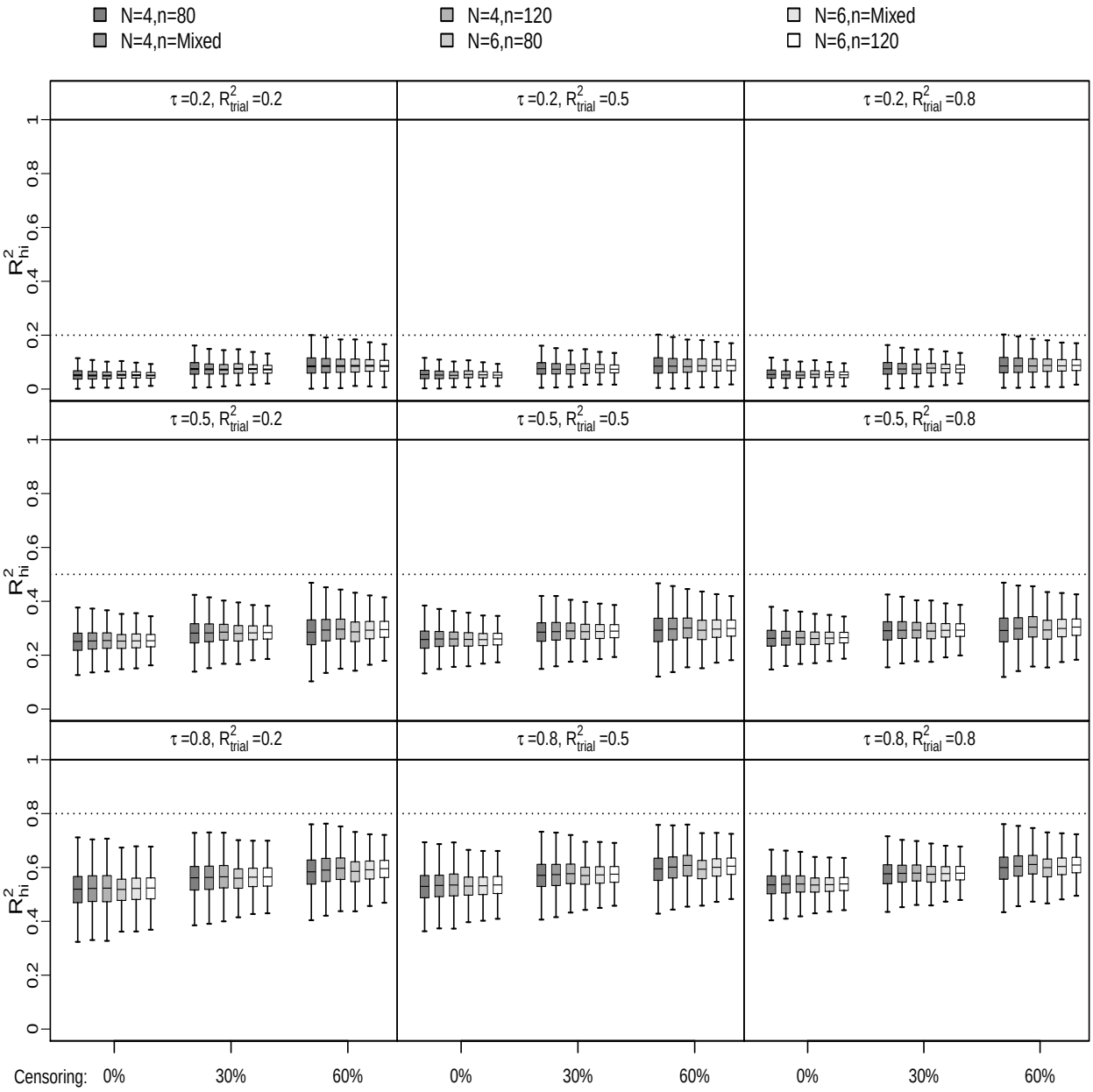

Figure 4.6: Boxplots of estimates of $R^2_{h,i}$: PFS, Clayton Copula Data Generation, Information Theory Application

Table 4.5: % Bias of Estimates of $R^2_{h,i}$ and $R^2_{h,t}$: $N = 6$, $n = 120$, PFS with Clayton Data

τ	R^2_{trial}	% Censoring	Median % bias $R^2_{h,i}$	$R^2_{h,t}$
0.2	0.2	0%	-74.694	144.127
0.2	0.5	0%	-74.161	27.848
0.2	0.8	0%	-73.650	-6.378
0.2	0.2	30%	-63.644	141.623
0.2	0.5	30%	-63.127	21.988
0.2	0.8	30%	-62.717	-10.983
0.2	0.2	60%	-57.327	135.138
0.2	0.5	60%	-56.599	15.921
0.2	0.8	60%	-56.014	-19.731
0.5	0.2	0%	-49.276	150.994
0.5	0.5	0%	-47.943	37.677
0.5	0.8	0%	-47.030	4.953
0.5	0.2	30%	-43.223	155.004
0.5	0.5	30%	-42.130	35.527
0.5	0.8	30%	-41.406	-0.638
0.5	0.2	60%	-40.996	156.077
0.5	0.5	60%	-40.176	25.556
0.5	0.8	60%	-39.132	-10.450
0.8	0.2	0%	-34.575	130.150
0.8	0.5	0%	-33.110	36.726
0.8	0.8	0%	-32.648	8.802
0.8	0.2	30%	-29.376	143.886
0.8	0.5	30%	-28.140	36.209
0.8	0.8	30%	-27.659	3.919
0.8	0.2	60%	-25.548	155.958
0.8	0.5	60%	-24.424	30.165
0.8	0.8	60%	-23.882	-5.231

interpretation. Again, it would appear difficult to conclude that a surrogate endpoint was a reliable predictor for OS, even when the true underlying association is strong.

Based on the results of TTP, and on the lack of dependence structure assumption within the information theory approach, it is expected that results of Gumbel data generation (Figure 4.7) would be similar to those from the Clayton generated data, with values being potentially slightly higher. Further, since there was some increase in values of $R^2_{h,i}$ when moving from TTP to PFS in the Clayton generated data, it is anticipated that the same pattern would occur for the PFS setting.

Figure 4.7 indeed shows that the pattern of increase in values of $R^2_{h,i}$ between data generation methods and surrogate endpoints is consistent, with values being slightly higher than both the PFS Clayton data and the TTP Gumbel data. The median percentage bias reduces by approximately 10% between the PFS Clayton and PFS Gumbel generated data under no censoring; a similar level to that observed under the TTP scenario. Under censoring, the reduction in bias is much greater (up to 30%), reflecting that increased censoring leads to higher estimates of $R^2_{h,i}$, which occurs for all strengths of association.

Further, the variability also increases as the proportion of censoring increases, particularly for low to medium association levels. This variability continues to hamper interpretation since there is very little to distinguish between truly medium and high levels of association. Whilst truly poor surrogates appear to be reliably identifiable, it is not possible to conclude that truly strong surrogates can be identified. Again, increasing sample sizes appears to reduce the variability marginally, but not sufficiently to conclude that the information theory method can detect promising surrogate endpoints.

Lognormal Data

As for the TTP setting, additional simulations were conducted to assess whether there was any impact on estimation through the use of a copula model to generate datasets. Given that results were broadly comparable across all three data generation algorithms for the

Table 4.6: % Bias of Estimates of $R^2_{h,i}$ and $R^2_{h,t}$: $N = 6$, $n = 120$, PFS with Gumbel Data

τ	R^2_{trial}	% Censoring	Median % bias $R^2_{h,i}$	$R^2_{h,t}$
0.2	0.2	0%	-78.845	122.174
0.2	0.5	0%	-78.438	20.046
0.2	0.8	0%	-78.086	-12.192
0.2	0.2	30%	-61.409	132.872
0.2	0.5	30%	-60.677	20.012
0.2	0.8	30%	-60.199	-13.982
0.2	0.2	60%	-38.759	130.115
0.2	0.5	60%	-38.231	12.594
0.2	0.8	60%	-37.683	-19.833
0.5	0.2	0%	-46.736	136.205
0.5	0.5	0%	-44.948	29.746
0.5	0.8	0%	-43.289	-2.761
0.5	0.2	30%	-31.258	152.665
0.5	0.5	30%	-29.428	30.428
0.5	0.8	30%	-28.247	-5.278
0.5	0.2	60%	-11.555	151.035
0.5	0.5	60%	-10.083	21.245
0.5	0.8	60%	-7.982	-11.285
0.8	0.2	0%	-22.998	125.080
0.8	0.5	0%	-20.012	32.275
0.8	0.8	0%	-17.930	4.564
0.8	0.2	30%	-14.752	139.266
0.8	0.5	30%	-12.381	33.562
0.8	0.8	30%	-10.674	2.245
0.8	0.2	60%	-6.470	151.237
0.8	0.5	60%	-4.447	29.312
0.8	0.8	60%	-3.137	-4.477

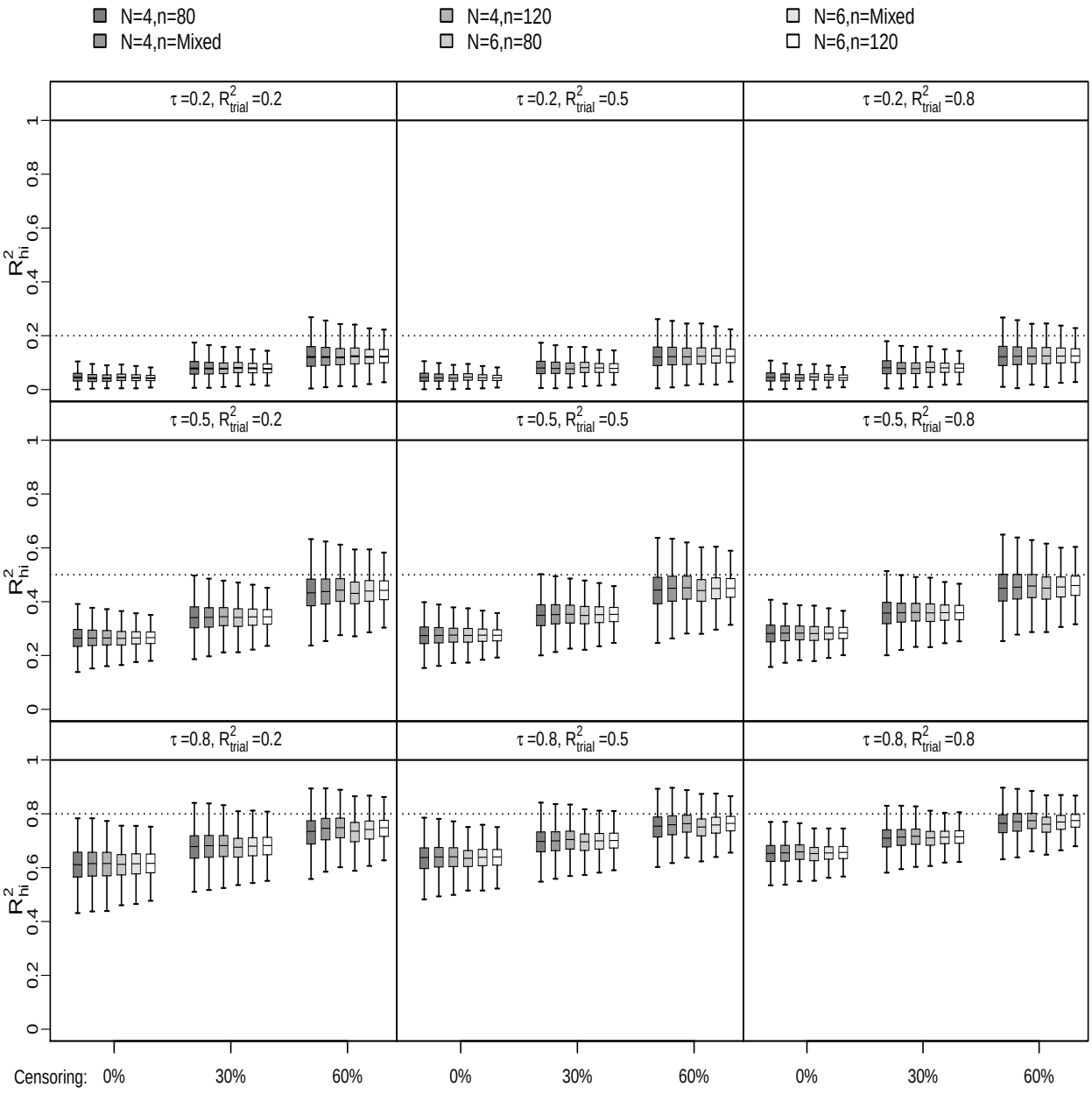

Figure 4.7: Boxplots of estimates of $R^2_{h,i}$: PFS, Gumbel Copula Data Generation, Information Theory Application

investigation based on TTP as the surrogate endpoint, it is expected that the same would be true when using PFS. Figure 4.8 presents results for the lognormal data (white boxes), alongside Clayton (light blue) and Gumbel (dark blue) for easier comparison. Horizontal dashed lines represent the true value of τ used in data generation.

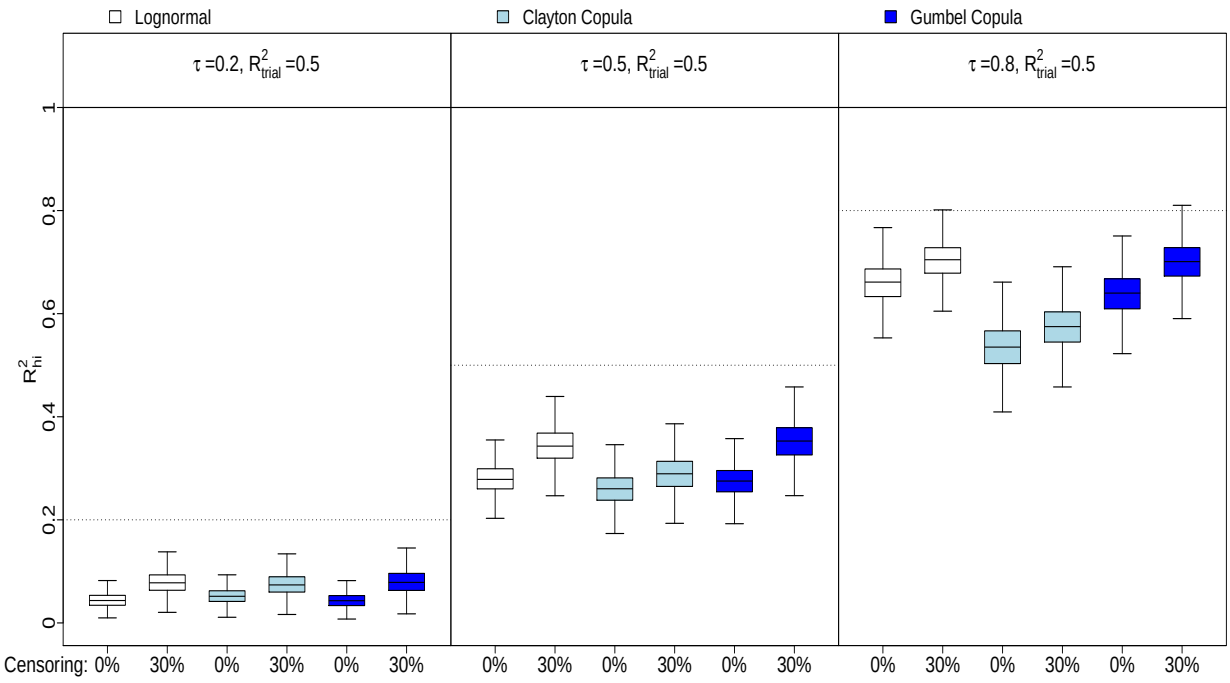

Figure 4.8: Boxplots of estimates of $R^2_{h,i}$: PFS, Information Theory Application to All Data Generation Methods (N=6, n=120)

The estimates of $R^2_{h,i}$ based on data generated using the lognormal approach are broadly comparable with those using the copula data. Across all settings, the estimates were higher for the lognormal than for Clayton generated data, but were consistent with those from the Gumbel generated data, and conclusions would remain consistent regardless of the data generation procedure. The pattern of increased results under censoring remained present, as did the large variability. Although results show that there is no overlap between medium and high levels of association, these results are based on the largest sample sizes of six trials each containing 120 patients. Results therefore demonstrate further consistency, between all three data generation algorithms, and between surrogate endpoints of TTP and PFS.

Overall, the results for PFS are consistent with what was observed for the TTP scenarios, and show that the information theory method is reasonably robust to both different data structures and surrogate endpoints that also incorporate data from the true clinical

endpoint. This is a key finding, since this provides more confidence in these settings where the two-stage meta-analytic copula method provided unreliable results. However, the wide ranges of estimates observed for the scenarios investigated here mean it is very unlikely that reliable conclusions can be made for truly strong surrogates. Further discussion of the comparison between the information theory and two-stage meta-analytic copula method is provided in Section 4.5.1.

Estimation of $R^2_{h,t}$

As for the TTP setting, there is very little difference in the estimates of $R^2_{h,t}$ between Clayton and Gumbel generated data, and so discussion of the results will be given for the two settings together. Results can be found in Figure 4.9 and 4.10, with supporting data presented in Tables 4.5 and 4.6 respectively.

Consistent with the two-stage meta-analytic copula method described in Chapter 3, estimates of trial-level association appear generally higher for the scenarios based on PFS than those for TTP. Where there was severe over- or under-estimation observed for the TTP scenarios, estimates of $R^2_{h,t}$ based on PFS as the surrogate endpoint are mostly over-estimated and rarely under-estimated, except where $R^2_{trial} = 0.8$. Estimates range across the entire $[0, 1]$ interval, with median values generally lying in the range of 0.5 to 0.8, and never lower than 0.4 even for the lowest true association level.

Whilst variability decreases slightly when the true association is at its highest ($R^2_{trial} = 0.8$), there is little that can be concluded due to the wide range of results. Again, improvement is observed when the number of trials increases from four to six, but this still does not allow for meaningful conclusions.

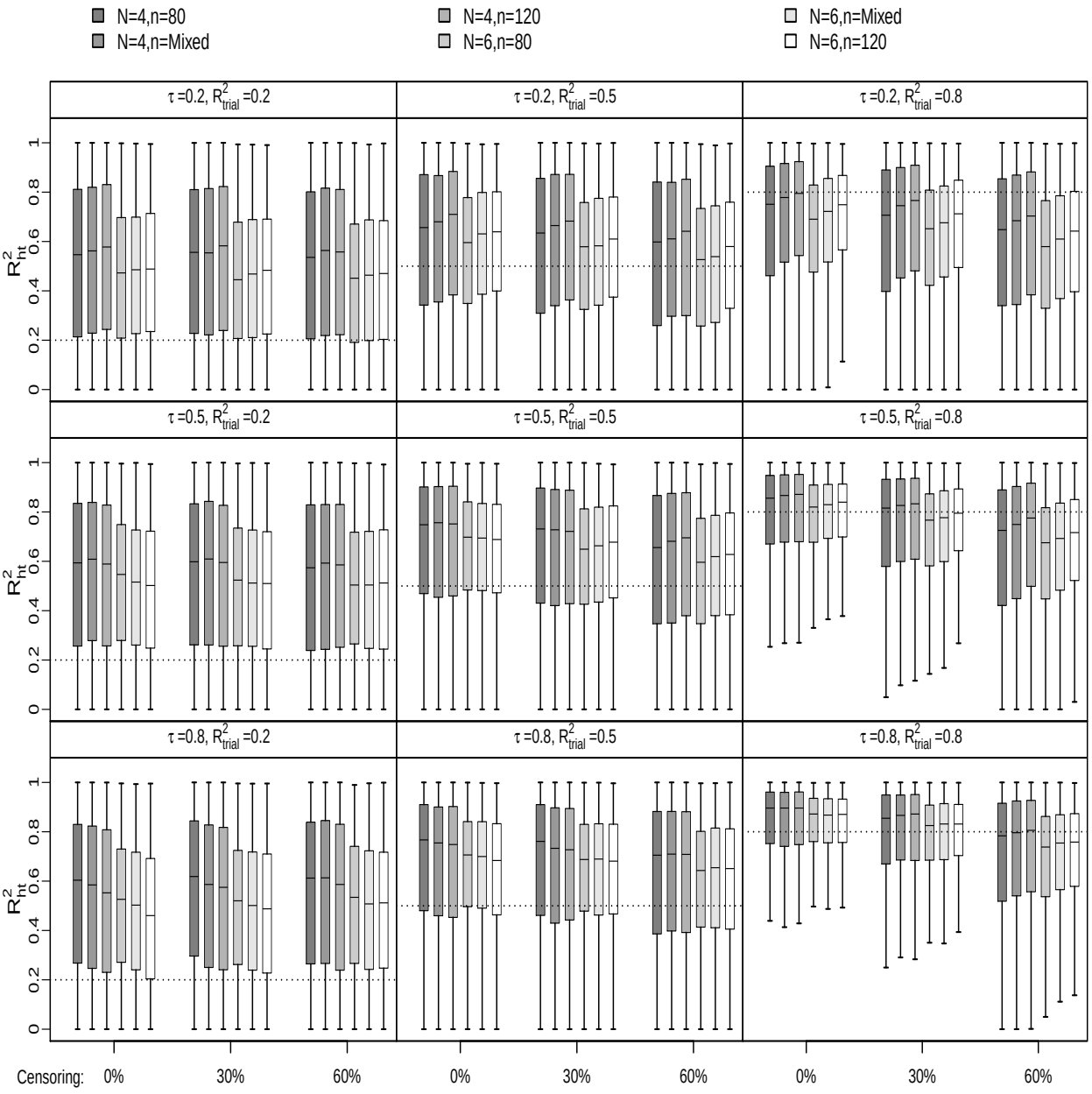

Figure 4.9: Boxplots of estimates of $R^2_{h,t}$: PFS, Clayton Copula Data Generation, Information Theory Application

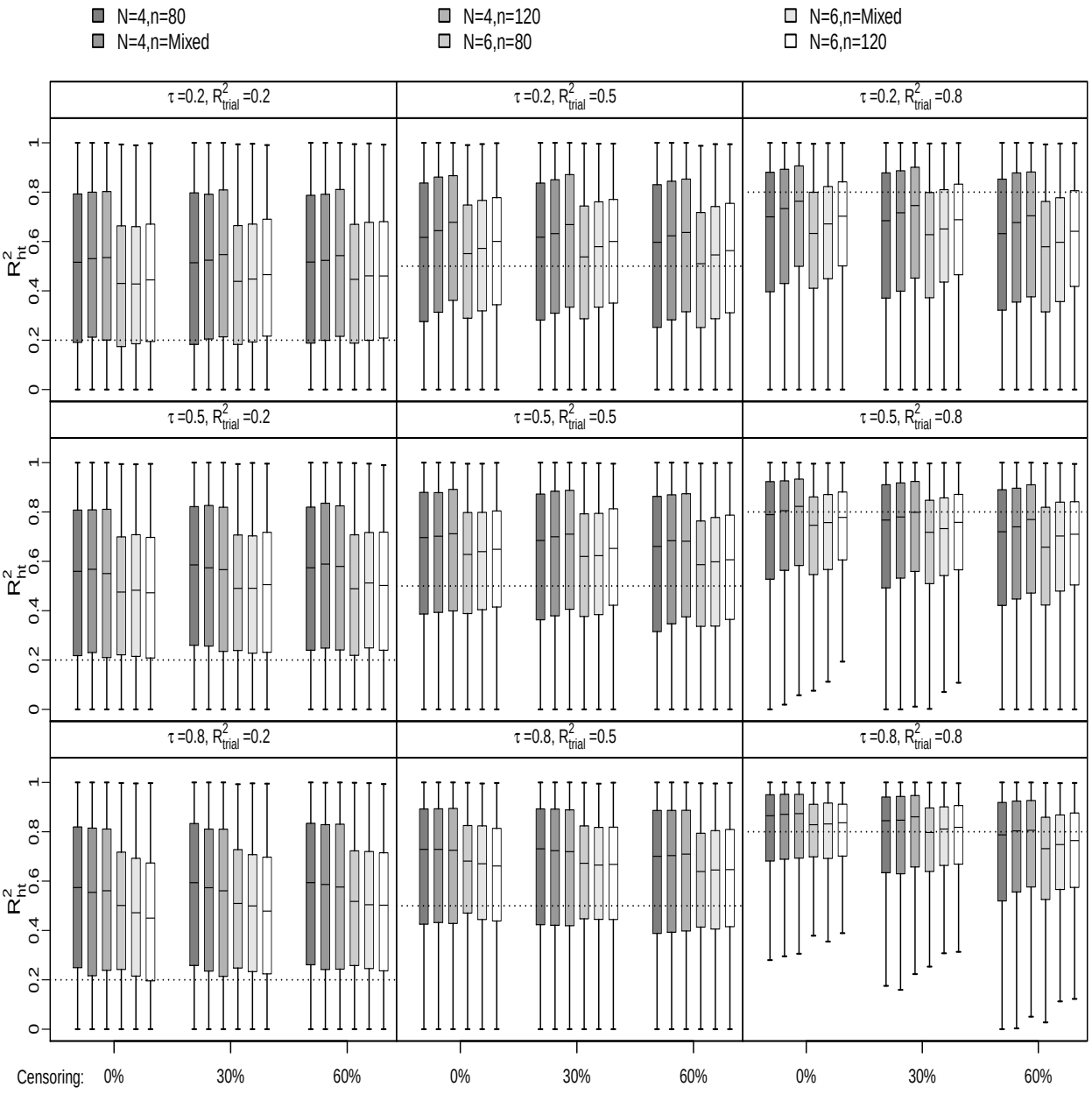

Figure 4.10: Boxplots of estimates of $R^2_{h,t}$: PFS, Gumbel Copula Data Generation, Information Theory Application

4.3.3 Further Exploration of Time Ordered Endpoints

Pryseley et al. (2011) note that the information theory model can be modified for exploration of time-ordered endpoints where $S \leq T$, such as the exploration of surrogates for overall survival. They suggest to replace the outcome of the proportional hazards models with the time difference between the surrogate and true outcomes, rather than the true endpoint value itself. This suggestion was made as part of the discussion of their findings, with a note that the computations and measure of association remain unaffected by this change. However, no examination of this alternative approach is made by Pryseley et al. (2011), and no further investigation of this suggestion appears to have been conducted in the literature. Since the ordering of endpoints appeared to cause significant issues in the performance of the two-stage meta-analytic copula method, having an approach that can appropriately handle such situations would be of great benefit.

To investigate further, the alternative modelling approach was explored for all scenarios under consideration in this simulation study. The outcome variable was taken 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. In this setting, there is no need to consider a time-dependent covariate to represent the surrogate, since the time period being modelled is that occurring after observation (or censoring) of S.

Across all scenarios, it was unfortunately found that the proposed adjustment to the model leads to extremely poor estimation of both $R^2_{h,i}$ and $R^2_{h,t}$. Changing the endpoint from TTP to PFS, and altering between Clayton and Gumbel-generated data had no impact on the results, with values of $R^2_{h,i}$ rarely exceeding a value of 0.2 even for the largest values of τ. Results can be found in Appendix B, Figures B.11 to B.18.

The most likely cause of the poor performance is that modelling of post-progression survival $(T - S)$ ignores the entire interval of time during which the patient remained alive and progression free, $[0, S)$. Since the modelling is based only on the time after the surrogate outcome has occurred, information on the duration of this prior period is

completely lost. The design of the simulation study aimed to achieve median times for PFS and OS of approximately $5-6$ and $10-11$ months, respectively, and so removal of the first portion of data leads to a substantial loss of information relevant to the surrogacy relationship.

When the true level of association is low, it could be expected that the impact could be the lowest, since knowledge of the surrogate time and outcome does not provide much predictive information on the true endpoint. However, increasing the strength of association between endpoints such that knowledge of the surrogate becomes highly predictive of the true outcome could be expected to have a substantial effect. The information captured by the surrogate includes not only the disease status but also the time at which the disease status changed. The modelling of $T-S$ does not reflect the time of disease progression, only the duration of time after progression that a patient survived, and so removal of this key information could be expected to weaken the relationship between disease status and true outcome, and therefore reduce the strength of surrogacy.

The overall implication of these results is that the alternative proposal for handling time-ordered endpoints can not be considered worthy of use. The information theory approach is not based on joint modelling of endpoints and does not assume any endpoint symmetry. The attempt to correct for a problem that did not exist with this modelling approach demonstrated that results were vastly inferior. 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.

Summary of Results

Results of investigation of the information theory approach lead to the following conclusions:

- The method is easy to implement and suffers minimal convergence issues.
- The sensitivity analysis using data generated from a lognormal model demonstrates that results based on copula generated data are interpretable and broadly robust.
- The highly consistent pattern of results observed between different surrogate endpoints and various data generation algorithms provides confidence that the method is robust to changes in dependence structure and symmetry of surrogate and true endpoints.
- Whilst there is no true reference value against which to compare estimates of individual-level association, increasing τ leads to larger estimates of $R^2_{h,i}$ which reflects the stronger relationship between endpoints.
- Large variability, particularly for medium to high association, limits interpretation and makes it difficult for the method to reliably identify good surrogate endpoints.
- There is evidence to suggest that the proportion of censoring in the data may affect estimation, with values increasing as the censoring proportion increases.
- Adjusting for time-ordered endpoints, such as modelling post-progression survival, does not provide reliable results and cannot be recommended.
- Estimation of $R^2_{h,t}$ based on only four or six trials is poor and cannot be recommended.

4.4 Understanding the Results

4.4.1 Comparison to Previous Simulation Study

The only known simulation study of the information theory approach to assessing time-to-event surrogate and true endpoints was conducted by Pryseley et al. (2011). Their study examined a number of approaches designed to estimate $R^2_{h,i}$, concluding that the estimation procedure of Xu and O'Quigley (1999), R^2_{XOQ}, could be recommended for future use. This estimation method was therefore used in the current simulation study, as described in Section 2.4.2.

Pryseley et al. (2011) generate TTP (as S) and OS (as T) data using the Clayton copula model, according to the same procedure as that used in the current study, but without controlling the trial-level surrogacy. All other steps of the data generation were identical between the two studies, aside from the range of simulation parameters being explored, the selected treatment effects on S and T and the chosen median survival times. Based on their 500 replicates of each simulation scenario, Pryseley et al. (2011) concluded that the method performs well, with slight under-estimation when the true τ is high (0.9), and absolute bias of $< 10\%$ for $\tau = 0.3$, 0.5 and $< 20\%$ for $\tau = 0.9$. There was little change through an increase in sample size, although this did improve estimation when the censoring was high (50%). When there was a high proportion of censoring on T, such that the T-dependent censoring of S (from the TTP setup) was $\geq 40\%$, the method was found to perform poorly, exhibiting under-estimation of high τ and over-estimation of low-medium τ.

The downward bias observed in the study of Pryseley et al. (2011) is consistent with the current study, however the bias values are substantially smaller. For the highest sample sizes, with $\tau \leq 0.5$ absolute bias in the current study reached as high as 84%, and for $\tau = 0.8$ as high as 55%, compared to $< 10\%$ ($\tau = 0.3$, 0.5) and $< 20\%$ ($\tau = 0.9$) from the previous study. To explore the reasons for this, a number of steps were taken. Firstly,

the R-program used by Pryseley et al. (2011) was obtained via a request to the author, and applied unedited to determine whether the results could be replicated. Following this, the code was closely examined to identify any differences in the modelling approach, and a number of issues were identified. Each of these will be described, with the aim of understanding their impact on the current study.

Replicating Results of Pryseley et al. (2011)

Results from the direct application of the unedited R-program provided by Pryseley et al. (2011) are provided in Table 4.7. All simulation parameters remain unchanged from this unedited program. As demonstrated in Table 4.7, bias values are very similar to those reported by Pryseley et al. (2011), suggesting that the code can reliably reproduce the published results and thus can be used as a basis to further investigate the differences in observed results between the two simulation studies.

Table 4.7: Results Using Code of Pryseley et al. (2011)

		Median % bias	
τ	% Censoring	Pryseley et al. (2011)	Re-Run
0.3	0%	−2.7%	−3.0%
0.3	20%	−3.0%	−3.0%
0.5	0%	−1.2%	1.6%
0.5	20%	−1.0%	1.0%
0.9	0%	−18.8%	−22.4%
0.9	20%	−18.7%	−22.7%

Simulation Procedure

The first discrepancy noted in the code used by Pryseley et al. (2011) is in the Clayton copula simulation. According to Equation (3.1) and the published article of Pryseley et al. (2011), the initially generated Uniform variables, U_{ij} and V_{ij} are transformed to be associated with strength θ_c and dependence structure of the Clayton copula through the

transformation

$$V_{ij} = \left(U_{ij}^{1-\theta_c} V_{ij}^{\theta_c^{-1}-1} - U_{ij}^{1-\theta_c} + 1 \right)^{\frac{1}{1-\theta_c}},$$

however the code from Pryseley et al. (2011) has a change in sign, using

$$V_{ij} = \left(U_{ij}^{1-\theta_c} V_{ij}^{\theta_c^{-1}-1} + U_{ij}^{1-\theta_c} + 1 \right)^{\frac{1}{1-\theta_c}}.$$

The data are therefore not generated according to the intended strength of association, and the value of τ, used as reference for the bias calculations, is subsequently incorrect. Updating the code to correct this sign leads to lower estimates of $R^2_{h,i}$ and therefore larger bias, as shown in Table 4.8.

Table 4.8: Results Using Code of Pryseley et al. (2011) with Corrected Sign

τ	% Censoring	Median % bias
0.3	0%	−78.7%
0.3	20%	−79.1%
0.5	0%	−67.2%
0.5	20%	−68.4%
0.9	0%	−28.7%
0.9	20%	−28.6%

These new bias values are closer to those observed in the current study, acknowledging the small differences in τ and the censoring proportions, as well as the variability that is likely present when only 500 simulations are run. Whilst the bias values for $\tau = 0.9$ are lower than those for the value of $\tau = 0.8$ observed in Table 4.3, it can be seen that as τ increases, the bias values are reducing, hence it is expected that this pattern would continue for higher values of association. These findings confirm that the information theory method demonstrates under-estimation (of τ) and suggests that the method may not be as appropriate as previously thought.

Although the study of Pryseley et al. (2011) is the only known investigation of the information theory approach in assessing surrogacy, the underlying measure of association, R^2_{XOQ} has previously been studied in a small simulation study of 100 runs (Xu and

O'Quigley, 1999). Since the context of this earlier study is not surrogacy, the data generation procedure used is simpler, with data being simulated according to the Cox regression model with a chosen strength of covariate coefficient. Although the covariate coefficient is not the eventual measure of association, selection of weaker ($\beta = \log(1)$), medium ($\beta = \log(2)$) and higher ($\beta = \log(64)$) model coefficients allows a general assessment of whether the measure R^2_{XOQ} is increasing with increased predictive value of the covariate. Results of this study showed R^2_{XOQ} values of approximately zero for $\beta = \log(1)$, of 0.1 for $\beta = \log(2)$ and in the range of $0.68 - 0.86$, depending on censoring, for $\beta = \log(64)$. Despite the limitations of the small number of simulation runs and differences in data generation, these results would suggest that a hazard ratio of approximately 2 could be expected to achieve a value of R^2_{XOQ} in the region of 0.1.

In order to investigate further, estimation of the β values and hazard ratios for TTP Clayton data generation are presented in Figure 4.11 for a variety of underlying values of τ $(0.1 - 0.9)$, based on six trials each containing 120 patients. On the left are the model coefficient values, β, and on the right are the corresponding hazard ratios, $\exp(\beta)$.

These estimates suggest that a hazard ratio approximately equal to a value of 2 could reflect a true underlying τ of close to 0.3, and so this level of individual association between TTP and OS, as explored by Pryseley et al. (2011), could have been expected to lead to estimates of $R^2_{h,i}$ close to 0.1. However, Pryseley et al. (2011) reported bias values for $R^2_{h,i}$ of approximately -3% for $\tau = 0.3$, suggesting that estimates were closer to a value of approximately 0.3. This further supports that the findings of Pryseley et al. (2011) were adversely affected by the apparent error in the simulation code.

Modelling Time-to-Event Outcomes

As part of the evaluation of code from Pryseley et al. (2011), another difference was discovered, relating to how patient risk-sets are defined. As per Equation 2.5, the information theory method is based on estimation of the conditional distributions of $T|S, Z$ and $T|Z$

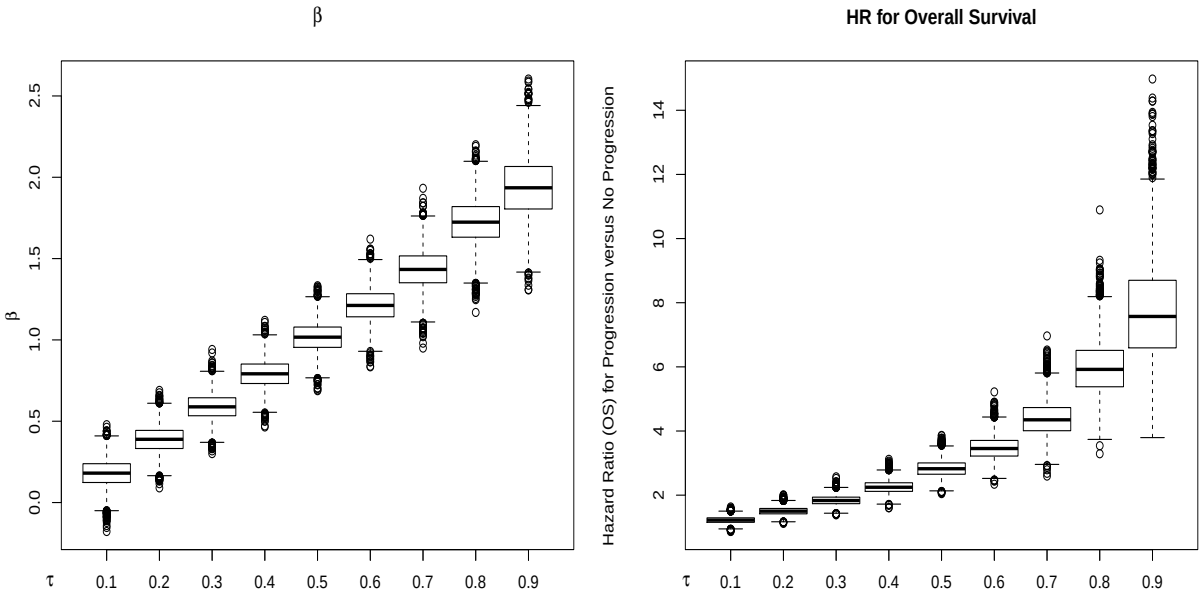

Figure 4.11: Boxplots of estimates of β and HR for OS: TTP, Clayton Copula Data Generation

at each event time, t_k. This requires calculation of the conditional probability of patient j having an event at time t_k, given their covariate values at that time:

$$\pi_j(t_k, \beta) = \frac{Y_j(t_k)\exp(\beta Z_j(t_k))}{\sum_{l=1}^{n} Y_l(t_k)\exp(\beta Z_l(t_k))}, \tag{4.3}$$

where the risk-set $Y_j(t_k)$ denotes an indicator variable to determine whether patient j is at risk of an event at time t_k (similar for patient l in the denominator) and $Z_j(t_k)$ denotes the covariate values for patient j (similar for patient l in the denominator) at that time.

Given that a time-dependent covariate is used to represent the surrogate outcome, the Z_j used in this calculation must reflect the covariate values that exist at each time t_k. The study described in this chapter therefore uses the covariate values from the interval $[0, S)$ when $t_k < S$ and covariate values from the interval $[S, T)$ once $t_k \geq S$, to ensure that the correct surrogate outcome is used at each t_k. In contrast, the code of Pryseley et al. (2011) considers patients at risk at each event time whenever t_k is lower than the upper end of the respective time interval ($[0, S)$ or $[S, T)$), leading to patients contributing twice

to the likelihood function for all t_k that are smaller than both S and T. Such a modelling approach is not felt to be appropriate, since the surrogate status changes between these two intervals. When $t_k < S$, the covariate values from the interval $[S, T)$ are not applicable, since this would be looking into the future; at time t_k the only covariate values should be those that are present just prior to that time. The impact of this is that the likelihood contributions are duplicated for these patients when calculating the probability of an event, biasing the estimates of $\pi_{ij}(t_k, \beta)$ and potentially leading to incorrect results. Since the approach taken by Pryseley et al. (2011) was not felt to be appropriate, no further investigation of this was conducted.

This investigation of the Pryseley et al. (2011) simulation study has demonstrated that results of the study may be misleading, with the error in simulation meaning that the value of τ used to estimate the bias is not accurately reflected in the data. Correction of this error led to bias values that are comparable with the findings of Section 4.3.1. It should be noted that further examination of the two alternative measures investigated by Pryseley et al. (2011) has not been conducted, and may be a topic for future research.

4.4.2 Underestimation

The downward bias in surrogacy estimates observed in the simulation study presented in this chapter occurs across almost all scenarios explored, particularly for TTP and for Clayton generated data. Results of PFS were slightly higher than for TTP, which could be expected since the former considers information from the true endpoint through the inclusion of death events in the definition. This additional information likely increases the strength of association between PFS and OS, thereby improving the under-estimation observed for TTP.

Results from Gumbel generated data were slightly higher than for Clayton data, and this is likely due to the difference in dependence structures induced by these two models. Where the Clayton copula demonstrates stronger late tail dependence, the Gumbel copula

exhibits more early tail dependence of event times. This early tail dependence suggests that it may be easier to predict which patients will be most likely to experience the OS event, therefore leading to a stronger relationship between the outcome and the covariate representing the surrogate endpoint. A subsequent increase in the estimates of the surrogate covariate coefficient was observed for the Gumbel data as compared to the Clayton data generation, thus leading to the higher values of $R^2_{h,i}$. The impact of censoring is considered to magnify this effect, with loss of the later, weakly associated event times leading to even stronger values for the covariate coefficient and therefore an increase in $R^2_{h,i}$ under censoring for the Gumbel generated data. The impact of censoring was substantially lower for the Clayton generated data, suggesting that there is minimal effect when event times demonstrate strong late-tail dependence.

Despite these particular cases where estimates are slightly higher, $R^2_{h,i}$ values continue to be lower than expected in the majority of scenarios. One potential reason for this is that the measure is known to be bounded by a number less than one when the covariates are discrete (with few levels) and the true association between outcome and a given covariate is very high (i.e. as $\beta \to \infty$). This is recognised by Pryseley et al. (2011), however Xu and O'Quigley (1999) note that this bound does not usually require 'special attention', and O'Quigley (2008) consider that it can be 'practically ignored'. Whilst this may be suitable when using the measure to generate prognostic models, the context of surrogacy requires high estimates of $R^2_{h,i}$ to demonstrate that the evidence supporting a surrogate endpoint is overwhelming. Results from the simulation study presented in this thesis show that it therefore may not be appropriate to disregard the issue of boundedness, since it may preclude observation of the levels of surrogacy that would be considered necessary for a surrogate endpoint to be considered reliable for future use. However, whilst this boundedness may contribute to the lack of observation of very high values of $R^2_{h,i}$, it does not help to explain why the lower levels of association are also estimated at values below those used in data generation. Overall, these issues demonstrate that the information theory approach has limitations in the setting of surrogacy of time-to-event outcomes.

Increased levels of association

Within the scenarios explored in this simulation study, the true underlying levels of individual association were restricted to 0.2, 0.5 or 0.8. Across all scenarios, results demonstrated that the estimated $R^2_{h,i}$ values are increasing with true underlying association, but not to the same degree as the true association value, particularly for TTP. It is not possible to see from the range of true τ values explored whether the level of association estimated by the information theory method reaches a plateau, or what strength of true association, if any, would result in truly high estimates of association (> 0.8). As such, further simulations based on TTP Clayton data generation were conducted with a wider range of τ values ($N = 6$ and $n = 120$, without censoring), to examine the full pattern of $R^2_{h,i}$. Results are presented in Figure 4.12.

As can be seen from this plot, high values of $R^2_{h,i}$ remain difficult to achieve, with maximum values reaching a value of only 0.66. However, at the highest level of $\tau = 0.9$, the Cox model coefficient for the surrogate endpoint, estimated as an intermediate step within the information theory method, reaches very extreme values (refer to Figure 4.11, left hand plot). Whilst such values would be a very good indication that a surrogate is highly predictive of long-term outcome, it is considered extremely unlikely to be achieved in practice, reflected by the large hazard ratios corresponding to these covariate coefficients (right hand side of Figure 4.11).

Although these additional results have therefore demonstrated that values of $R^2_{h,i}$ continue to increase as the true level of τ increases, the strength of surrogacy that would be required to attain high estimates of individual-level association become infeasible. Further exploration of PFS was not conducted with higher τ values, since the treatment effect here is expected to be even more extreme due to the composite nature of the PFS endpoint. From these additional simulations, it is concluded that whilst the information theory method can theoretically estimate very high levels of prediction, this would appear to be rarely feasible in practice.

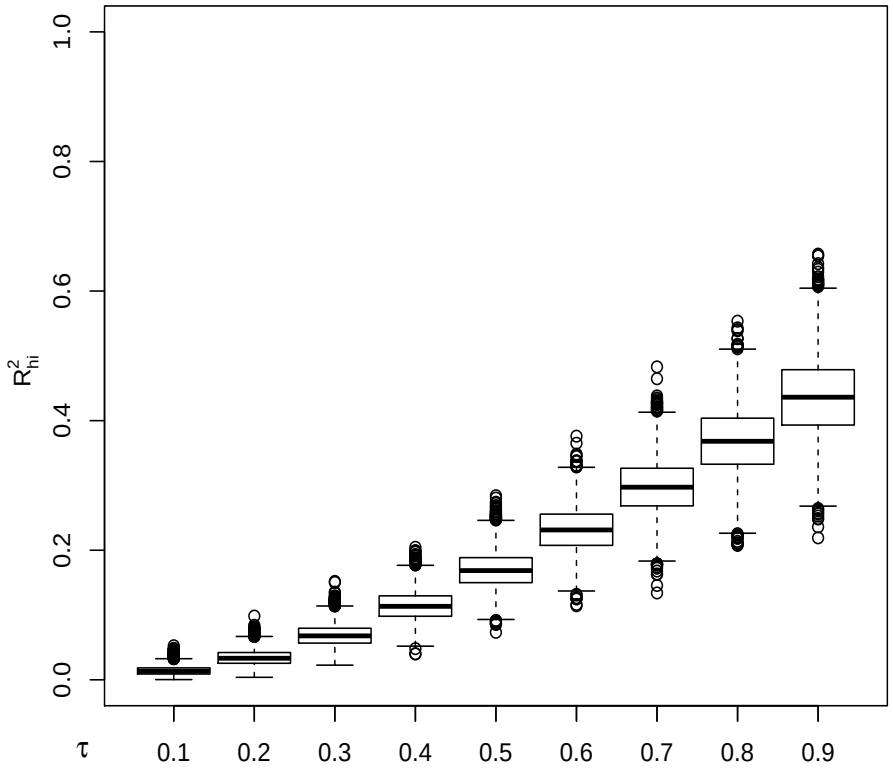

Figure 4.12: Boxplots of estimates of $R^2_{h,i}$: TTP, Clayton Copula Data Generation, Information Theory Application

4.4.3 Variability

A striking feature of the simulation results presented in this chapter is that the variability in estimates of $R^2_{h,i}$ appears to increases as the strength of association increases. For the highest association level of $\tau = 0.8$, the results across all scenarios are highly variable, with the range stretching up to half of the unit interval. This is one of the major drawbacks of the approach, as it limits the interpretability of results and prevents reliable conclusions.

Further investigation into this issue has demonstrated that the increase in variability is likely due to the variability in estimation of Cox proportional hazards model coefficients,

particularly for the surrogate endpoint. Whilst the range of parameter estimates increases slightly as the strength of association increases, the effect of taking the exponential of this parameter, as is done when calculating the conditional probability of each subject having an event (Equation 4.3), means that the small increases in variability go on to have a large impact in subsequent estimation procedures. Since these exponentiated values are key in calculation of $R^2_{h,i}$, this leads to a much larger range of estimates of $R^2_{h,i}$, as evidenced in the simulation study results. The issue, therefore, is not directly with the information theory approach, but with the underlying modelling structure. This is demonstrated in Figure 4.11, where the modest increase in covariate coefficients (β, left hand side of the plot) leads to very large increases in variability when the hazard ratio ($\exp(\beta)$) is calculated (right hand side of the plot).

Therefore, as the strength of association increases, it is considered that there is increased uncertainty in the parameters of the Cox proportional hazards model, leading to increased variability in parameters that are subsequently used to estimate $R^2_{h,i}$. It is expected that increased sample sizes would improve the estimation of parameters, thereby reducing variability and leading to more reliable conclusions from $R^2_{h,i}$. For completeness, this aspect has been considered and discussion is provided in Section 4.4.4. Nevertheless, specifically for the sample sizes examined in this study, the wide variability reflects the uncertainty introduced through the lack of available data.

4.4.4 Larger Sample Sizes

Although the simulation study of Pryseley et al. (2011) considered larger sample sizes in their assessment of the information theory approach to evaluating surrogate endpoints, the issues described in Section 4.4.1 suggest that these results may not be entirely accurate, leading to no published literature on the performance of the measure under the ideal setting of larger sample sizes.

As a result, additional simulations were conducted for a selection of the scenarios

considered in the simulation study described in Section 4.2, with larger numbers of trials and patients within trials, to allow a more general interpretation of the performance of the surrogacy approach. A total of $5,000$ simulations were conducted using both TTP and PFS as surrogate endpoints for OS, with 0% and 30% censoring and with ten trials each containing 500 patients. To achieve such sample sizes would likely require a wide range of clinical trial datasets, potentially combining data from multiple molecules and different companies. Results of these additional simulations are presented in Figures 4.13 and 4.14 for TTP and PFS, respectively.

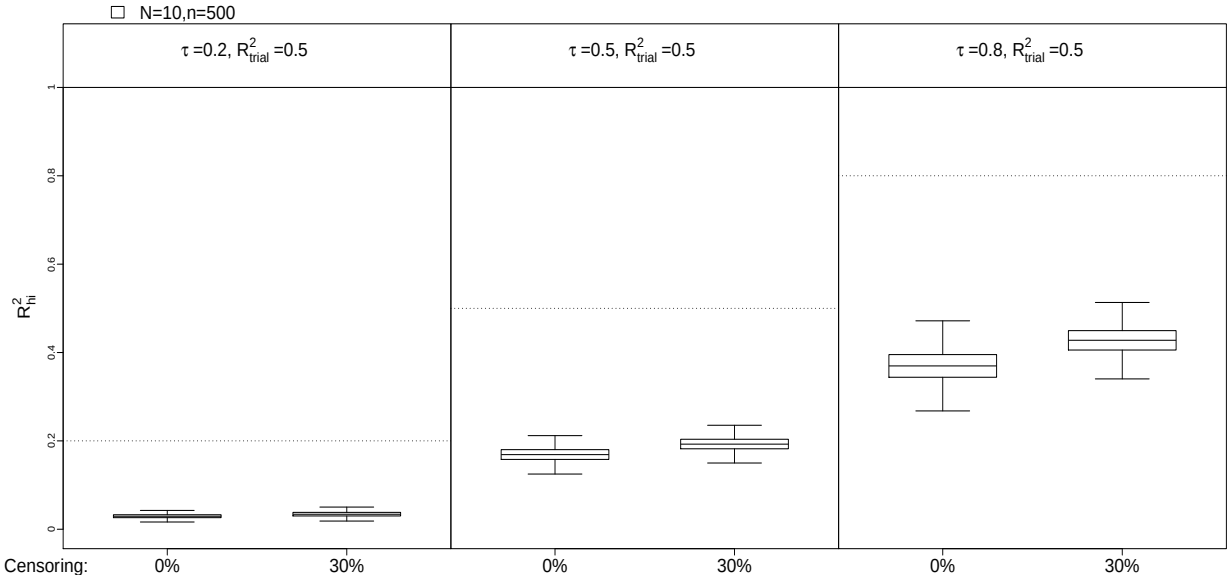

Figure 4.13: Boxplots of estimates of $R^2_{h,i}$: TTP, Clayton Copula Data Generation, Information Theory Application (larger sample sizes: $N = 10$, $n = 500$)

For both TTP and PFS scenarios, the availability of larger clinical trial databases has improved estimation, in particular with respect to the variability in estimated values of $R^2_{h,i}$. 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 no overlap of estimates between true underlying surrogacy values.

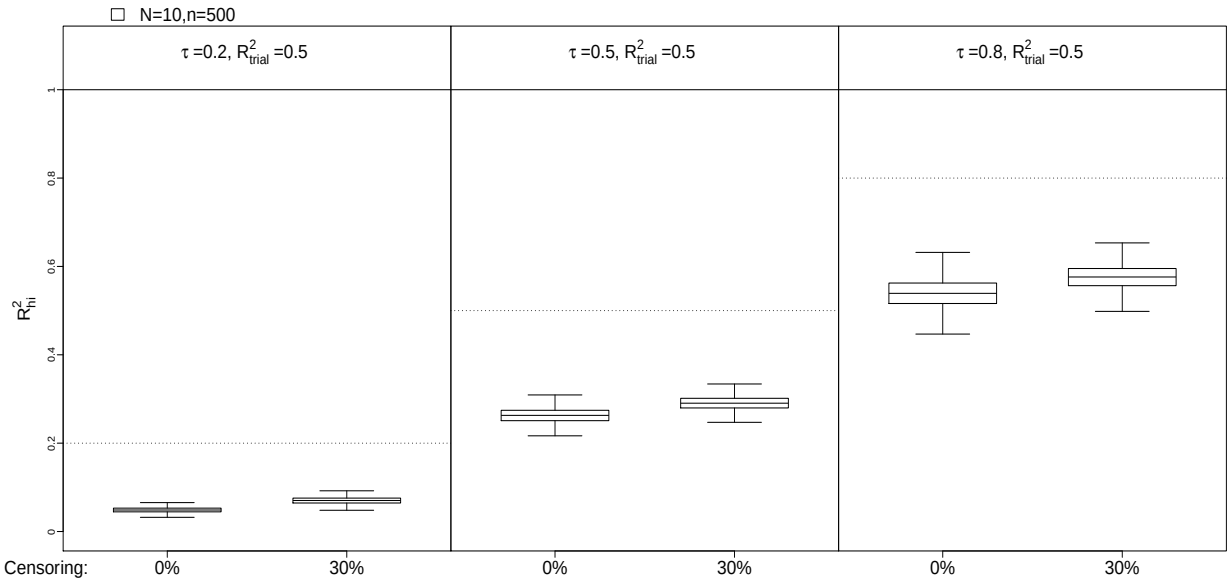

Figure 4.14: Boxplots of estimates of $R^2_{h,i}$: PFS, Clayton Copula Data Generation, Information Theory Application (larger sample sizes: $N = 10$, $n = 500$)

However, it remains evident that even under truly strong association, the method cannot reach values of $R^2_{h,i}$ greater than approximately 0.5 for TTP and 0.7 for PFS. The drawback to the reduced variability is reflected in the tighter ranges of estimates, which have led to a reduction in the maximum value of $R^2_{h,i}$ achieved during simulation. Therefore, while improved variability can be achieved through increased data, the underestimation remains of concern and promising surrogates may be overlooked.

4.5 Implications of Results

4.5.1 Comparison to Two-Stage Meta-Analytic Copula Method

The simulation studies and deeper investigation of the information theory approach described in this chapter and of the two-stage meta-analytic copula method described in Chapter 3 examined the performance of the two surrogacy evaluation approaches for the same scenarios and using identical datasets. Although these two methods are estimating different quantities, it is important to understand how consistent the conclusions of sur-

rogacy would be between the two. Despite the potential bias in direct comparison due to the use of copula functions in the data generation algorithms, both approaches were also applied to datasets that were not based on a copula function, and these sensitivity results demonstrated that there was no impact from the selected simulation algorithm (Figures 4.3, 4.8 for information theory, Appendix Figures A.9 and A.10 for the two-stage meta-analytic copula method). Therefore, a comparison of the performance of the two methods is considered appropriate.

As expected, when considering a surrogate endpoint of time-to-progression under perfect model specification of the Clayton copula, the two-stage meta-analytic copula method demonstrated superior performance in estimating individual-level surrogacy. This was evident across all scenarios. Despite high variability of the approach under some scenarios, the under-estimation of the information theory method means that even very promising surrogates would likely be rejected, limiting the interpretability of the method. Even when the two-stage meta-analytic copula method is subject to model misspecification such that performance deteriorates, it remains superior to the information theory method, which continues to demonstrate under-estimation to a level that does not allow for truly strong surrogates to be identified. Even with allowance for the parameter τ being different to that estimated by the information theory method, it appears difficult to achieve truly high estimates of individual-level surrogacy.

The arguably more impactful comparison arises when considering progression-free survival as a potential surrogate endpoint. Violation of the symmetry assumption of copula models led to significant deterioration in the two-stage meta-analytic copula method, and this is where the information theory method provides potentially superior performance. Despite the large variability in the results for the sample sizes investigated here, the information theory approach appears insensitive to both the underlying data dependence structure and the choice of surrogate endpoint. Violation of the symmetry assumption therefore has minimal impact, with changes reflecting what could reasonably be expected from the change in endpoint. This is considered a notable advantage, since the poten-

tially misleading results of the two-stage meta-analytic copula method could lead to very poor surrogates being used in practice in new Phase III trials, and potentially leading to regulatory approval of new treatments that ultimately will not offer benefit to patients in the most clinically relevant true endpoint. Whilst the under-estimation of the information theory approach is of concern, the consequences of making a decision of poor surrogacy are far less harmful.

In addition to these findings of relevance to the estimation of individual-level surrogacy, both surrogacy evaluation approaches were also used to estimate trial-level surrogacy. Uniformly across both methods and all simulation scenarios explored, performance was poor, and neither method can be recommended for use in estimating trial-level surrogacy when there exist data from only a small number of clinical trials each containing a small number of patients.

4.5.2 Practical Implications

As has been previously described, the non-convergence rate observed in the simulation study of the information theory method was lower than 1% across all simulation scenarios. This reflects the ease of computation, helped by the fact that the measure is based on quantities that are estimated by standard software packages. The software procedures that are needed to estimate the model parameters are those well known to researchers who work in the analysis of time-to-event endpoints, and this allows quick understanding of the underlying concept of the approach.

Further, the lack of assumptions around the joint distribution of endpoints means that the information theory approach can be considered applicable to a wide range of datasets without the need to change the modelling structure. This is an additional advantage, and again improves the ease of use. The only assumption that needs to be satisfied is that the data demonstrate proportional hazards, such that the Cox proportional hazards model is appropriate for use and resulting model parameter estimates are robust. Any

deviations from this assumption need to be taken into account, for example through the use of multiple time dependent covariates, potentially including treatment.

Whilst the potential benefits of the information theory method are clear, results of the simulations have shown two areas of concern; the high variability in estimates and the potential under-estimation of the strength of association between endpoints. The level of variability observed in the results suggests that whilst truly poor surrogates can be reliably identified, there is substantial overlap in estimates of medium to high individual-level surrogacy that prevents a clear conclusion. This unfortunately hinders the use of the information theory method for the setting of interest in this thesis. The inability of the information theory approach to reach truly high values of individual-level surrogacy also prevents recommendation of the measure for use in practice. It is critical that any surrogacy measure can reliably predict which surrogates have strong association with the true endpoint. A τ value of 0.8 demonstrates a very strong relationship, however the information theory method could provide a surrogacy measure lower than 0.2 in this scenario. Whilst some variation from the true input value of τ should be expected, this result is greatly contradictory, and could result in the rejection of truly beneficial surrogate endpoints. Whilst vast improvements were seen when examining much larger sample sizes, the issue remains and would certainly be of concern for the majority of practical surrogacy evaluations.

4.5.3 Limitations of the Simulation Study

The main limitation of the simulation study described herein has been noted previously; the true underlying individual-level surrogacy cannot be controlled via simulation. Since the $R^2_{h,i}$ 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 consistent across all scenarios investigated, therefore the conclusions are considered to be robust. The results of the simulation study clearly suggest that the information theory approach struggles to reach a level that would enable confidence in the strength of a given surrogate endpoint.

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). In the data generation procedure, a change in treatment effect has no impact on the underlying strength of association, τ, 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 (HR ≈ 0.50 for PFS and ≈ 0.67 for OS) based on TTP using Clayton copula data generation. The range of simulation scenarios included τ of 0.2, 0.5 and 0.8, with R^2_{trial} fixed at a value of 0.5, under no censoring and moderate (30%) censoring. Results are presented in Appendix B, Figure B.10, and demonstrate findings that are highly consistent with the originally selected treatment effects, with the median and ranges of estimates of $R^2_{h,i}$ being very similar across all levels of τ. 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), who concluded that the measure $R^2_{h,i}$ performed acceptably well when the proportion of censoring was low to mod-

erate, 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.

4.6 Further Work

The extensive simulation study and subsequent investigation of the information theory method described within this chapter has led to the conclusion that whilst the approach is subject to limitations through the high variability and difficulty identifying truly high levels of association between endpoints, the underlying concept has potential. It is easy to understand, and the avoidance of the joint modelling required by the two-stage meta-analytic copula method makes the information theory approach very easy to implement in practice. Further examination of larger sample sizes has demonstrated that estimation can be improved when there exist large amounts of data, and assessment of alternative representations of PFS as a surrogate endpoint has allowed recommendation of the most appropriate approach.

A substantial amount of literature has been published in the field of dependence measures for survival outcomes, including studies comparing multiple proposed measures of association. Many of these measures are based on the same underlying approach as that used within the information theory method for assessing surrogacy, potentially providing alternative methodological approaches to estimate the strength of association between surrogate and true endpoints.

The next steps of the research described in this thesis are to explore these alternative proposals and consider whether any of these may provide interpretable measures in the context of surrogacy. Discussion of alternatives, as well as development of a new approach to the evaluation of surrogacy, will be described in the next chapter.

Chapter 5

A Novel Approach to Evaluating Time-to-Event Surrogate Endpoints

5.1 Introduction

Despite the identified limitations of the information theory approach to evaluating surrogate endpoints, the underlying concept has appeal; it is computationally simple, easy to understand and implement, and appears insensitive to the type of surrogate endpoint or dependence structure within the data. As a result, alternative measures of association based on similar models are worthy of consideration as potential candidates for surrogacy evaluation.

The information theory method is based upon a measure of explained randomness in proportional hazards models, and is a measure of how much of the 'randomness' in the survival outcome can be explained through the inclusion of covariates, including the potential surrogate endpoint(s). This is analogous to measures of explained variation in linear models, however extension to survival models is non-trivial due to the necessary incorporation of censored information, the need to incorporate a time-dependent surrogate outcome and the fact that, when used, the proportional hazards model does not have a distributional assumption. The topic of association measures for survival data is one that

has undergone much research, with many different approaches being proposed for use.

Within this chapter, alternative measures of association for time-to-event data are discussed, and their application to the context of surrogacy evaluation is assessed (Section 5.2). Following a review of the applicability of available methods, one particular measure proposed for use in creating prognostic models for patient subgroups is selected for further consideration (Section 5.3). In contrast to prognostic modelling, this measure is evaluated and applied in the new context of assessing surrogate endpoints, to determine whether it may offer improvements over those measures already investigated within this thesis (Section 5.4). Subsequently, an extension of the approach to improve reliability in surrogacy evaluation is described in Section 5.5, and investigated via a simulation study. Multiple sensitivity analyses conducted to critically examine the new approach are described in Section 5.6. Further discussion of the results can be found in Section 5.7, with implications discussed in Section 5.8 and suggestions for further work presented in Section 5.9.

5.2 Measures of Association for Time-to-Event Endpoints

Association measures are intended to reflect and quantify the strength of relationship between an outcome variable and a set of one or more covariates. In linear modelling, such measures are well defined and are commonly used, however the complexities inherent in survival data mean that extension to time-to-event endpoints is challenging. There have been a number of approaches proposed and tested, yet there is no overarching consensus as to which may be the most reliable for practical use.

In an attempt to address this, Choodari-Oskooei et al. (2012a,b) consider a total of 17 different measures that have been proposed for survival models to provide an estimate of how accurately covariates can predict a survival outcome. These in-depth explorations of methods focus on the prognostic value of patient characteristics, and in particular how these characteristics can be used to build prognostic models for particular diseases. The 17 methods are categorised into measures of explained variation, explained randomness

or predictive accuracy. Explained variation methods quantify the proportion of outcome variability that can be predicted through the covariates included in the model. Explained randomness methods, such as the information theory method described in Chapter 4, are based upon the information, or entropy, of a distribution and estimate the improved precision in prediction of survival outcomes based on having knowledge of given covariates. Predictive accuracy measures compare the survival status for an individual at a given time to the predicted survival probability from models with and without covariates, providing an estimate of how well the addition of covariates to the model can improve this prediction.

Many of these 'R^2-type' measures have an intuitive interpretation that could also be considered appropriate for an evaluation of surrogacy. Rather than building prognostic models based on patient demographic and disease characteristics, which is the current proposed use of the methods, they could be considered applicable to assessing whether a surrogate endpoint can reliably predict true long-term outcome, with the prognostic value of the surrogate endpoint being captured via use of a time-dependent covariate. Relevant findings of the studies conducted by Choodari-Oskooei et al. (2012a,b) are therefore described in subsequent sections, with the aim to assess which, if any, of the proposed measures could be potential candidates in a new context of surrogacy evaluation.

5.2.1 Performance of R^2 Measures Using Survival Data

Explained Variation

The study of Choodari-Oskooei et al. (2012a) examined the performance of five different measures from the explained variation category (Kent and O'Quigley, 1988; O'Quigley and Flandre, 1994; O'Quigley and Xu, 2001; Royston and Sauerbrei, 2004; Royston, 2006). This investigation was conducted using simulations, in particular investigating the impact of censoring, different covariate distributions and robustness against influential (extreme and outlier) observations. Whilst each of the five selected approaches was shown to have limitations, two of the proposed explained variation measures were recommended for use

based on their ease of understanding to non-statisticians and their satisfactory performance under the simulation scenarios investigated (Kent and O'Quigley, 1988; Royston and Sauerbrei, 2004). Use of both of these measures together is recommended, to ensure that the limitations of each approach are considered and accounted for. However, since neither of these two approaches is able to incorporate time-dependent covariates, they are not applicable in a surrogacy context and will not be discussed further here.

Explained Randomness

Subsequently, Choodari-Oskooei et al. (2012b) further examine the remaining categories of measures; those based on explained randomness or predictive accuracy. Investigations of explained randomness measures include the information gain approach of Xu and O'Quigley (1999) which is implemented within the information theory surrogacy evaluation method described in Chapter 4. The other investigated measures can be found in Kent and O'Quigley (1988) and O'Quigley et al. (2005). The authors assess various factors of interest such as the distribution of covariates, varied proportions of censoring and the impact of influential observations. Based on simulation studies, estimated values of explained randomness appeared to be higher than those of explained variation, however all of the explained randomness measures investigated were impacted by the distribution of the covariates and the presence of influential observations. Choodari-Oskooei et al. (2012b) suggest that two of the measures performed well, however these are either complex to calculate or may be impacted when the linear combination of covariates and model coefficients is skewed (Kent and O'Quigley, 1988).

Overall, it is concluded that all of the explained randomness measures investigated have shortcomings that prevent a universal recommendation as to which demonstrates the best performance. Interestingly, the information gain method of Xu and O'Quigley (1999) is found to be influenced by the covariate distribution and by influential observations in the data, preventing a recommendation for its use in practice. In fact, it is concluded

that only one of the measures under investigation (and approximations thereof) could be considered reliable enough for use (Kent and O'Quigley, 1988), and this method cannot incorporate time-dependent covariates, preventing the application of the approach within the surrogate context.

Predictive Accuracy

The final category of measures explored by Choodari-Oskooei et al. (2012b) are those which evaluate predictive accuracy. Rather than estimating how much of the variability or randomness in survival outcomes can be explained by the covariates, these measures assess how well the covariates can predict the overall survival status of individuals at a given time. Choodari-Oskooei et al. (2012b) conduct a simulation study of two predictive accuracy measures, concluding that the measures can depend on the size of the covariate effect and the covariate distribution (Graf et al., 1999; Schemper and Henderson, 2000). However, neither measure was found to be sensitive to influential observations and both performed satisfactorily under small to moderate censoring. Importantly, the observed values of predictive accuracy measures tended to be smaller than the explained variation or randomness measures, as they quantify the uncertainty in prediction of a binary survival status at a given time rather than the uncertainty in the actual survival time itself. The predictive accuracy measures were also found to depend on the follow-up period. Since survival status and survival probabilities change over time, an arbitrary timepoint must be selected at which predictive accuracy is estimated, and results showed that this can have an impact on the result, with the measures generally increasing as time increases.

These detailed studies of measures proposed to quantify how well covariates can predict outcome in survival models demonstrate that measures based on explained variation are considered the preferred option, but provide estimates that are lower than those based on explained randomness. Therefore, if considered in the surrogacy setting of interest in this research, such measures are unlikely to provide estimates higher than those from the

information theory method, and are unlikely to improve the under-estimation observed with this approach. Similarly, the two predictive accuracy measures were found to provide the lowest estimates of association across all categories explored, suggesting that these also may not offer improvements over the information theory approach. In addition, the two measures of explained variation and one measure of explained randomness that were considered to be the most reliable cannot incorporate time-dependent covariates, and are therefore unsuitable for use in evaluation of surrogate endpoints. These comparison studies therefore do not provide alternative approaches that could be considered superior to the information theory method if applied in a surrogacy setting.

Since these investigations, Choodari-Oskooei et al. (2015) have proposed an alternative approach which quantifies how much the prediction of survival status can be improved through consideration of covariates. This measure is based upon the difference between the predicted survival probability from a model with covariates, and the average survival probability for all patients in the sample without accounting for covariate values. The difference between these two values, across all patients in the data, then provides a summary of how much improvement in prediction of survival status can be gained through the inclusion of covariates in the model. Originally proposed for binary outcomes (Bura and Gastwirth, 2001), the measure is explored in a survival setting by Choodari-Oskooei et al. (2015) through the use of simulations, again assessing the impact of censoring, covariate distributions (including time-dependent covariates) and influential observations. It is concluded that the proposed measure performs well with regards to characteristics that a measure of predictive ability needs to demonstrate, and is recommended as a measure to quantify predictive ability in survival models. This measure, termed Total Gain, is therefore described next, with an aim to further develop the approach to be applicable in the evaluation of surrogate endpoints.

5.3 Total Gain

5.3.1 Background

Bura and Gastwirth (2001) introduced the concept of 'Total Gain' (TG) to quantify the explanatory power of a model in predicting binary outcomes. By comparing the predicted probability of a binary outcome after adjusting for covariates to that based only on the average probability for the sample, the approach provides a measure of how much better the prediction can be once the covariates are taken into account. In particular, Bura and Gastwirth (2001) display this graphically, to allow for a visual representation of the model as well as visual comparison between models which contain different sets of covariates.

TG is calculated using two quantities; (1) the average probability of the event of interest, unadjusted for covariates, and (2) the predicted probability of the event of interest from a model adjusted for covariates. These two probabilities are defined for each patient within a sample, with the former remaining constant for all patients, and the latter being defined by the combination of covariate values for each individual patient. The absolute difference in these two probability values, taken across all patients, provides estimation of how much more accurate the prediction of outcome can be when covariates are taken into account.

In the binary setting, the average probability is calculated as the mean of the response variable, and the predicted probability is derived from a logistic regression model which contains one or more covariates. Since such models can include an arbitrary number and selection of covariates, Bura and Gastwirth (2001) propose to calculate TG over the percentiles of the distribution of linear predictors, where the linear predictors are calculated as the linear combination of covariate values and respective covariate coefficients estimated by the model fitting process. Basing calculations on the percentiles of these predictors rather than the values themselves allows multiple models to be compared on the same scale.

Further description on the calculation of TG, with example graphics, are provided in the next section, after the measure has been described in the context of survival outcomes.

5.3.2 Application to Survival Data

The extension of the TG measure to survival outcomes was proposed by Choodari-Oskooei et al. (2015), who recognised that improvement in prediction of survival status could provide a measure of the importance of available demographic or prognostic information. This may be particularly important in oncology settings, where overall survival continues to be a primary measure of outcome and is often a question that arises from patients to their physicians at the time of diagnosis. The use of TG potentially allows the practical value of baseline characteristics to be assessed via a simple, easy to calculate approach, and the graphical representation improves the ease of understanding for non-statisticians.

As described in the previous section, calculation of TG is based on two probabilities. The first is the average probability of the outcome of interest for all patients in the sample, with no adjustment for covariates. This provides the most basic prediction and creates a baseline against which improvements would be sought. In the binary setting, the mean of the response variable is used. For the time-to-event case, the outcome of interest is the occurrence of a particular event (e.g. survival) at a given time, and so the Kaplan-Meier function is proposed by Choodari-Oskooei et al. (2015) as a suitable method to provide an overall estimate of the probability of remaining event-free. Such a method provides an estimate of the probability of remaining event-free for all patients in the sample, whilst appropriately accounting for patient censoring. The second quantity that needs to be estimated is the predicted probability from a model containing covariates. In the binary setting, this is represented by a logistic regression model, which allows for prediction of the probability of experiencing the event of interest based on given covariate values. Similarly, for the time-to-event setting, a Cox proportional hazards model can be used to provide the estimated probability of remaining event-free at a given time, based on a set

of (possibly time-dependent) covariates. One advantage of the TG approach is that any survival model could be used in this step, depending on the best representation of the data. Linear predictors are constructed from the model in the same way as for a logistic regression model, using the linear combination of covariate values and estimated covariate coefficients from the chosen model. These linear predictors are then scaled based on their percentiles for the calculation of Total Gain, to allow direct comparison between different models.

Given that time-to-event data incorporate a time element, each of these two predicted probabilities must be estimated at a given, fixed time, t. This issue will be discussed further in the forthcoming sections, and the notation $TG(t)$ will be used when describing the measure in the survival setting. To illustrate the graphical representation of $TG(t)$, a theoretical example is provided in Figure 5.1, based on a single continuous covariate. In this graphic, the predicted probabilities of remaining event-free at the selected time, t, based on the Cox proportional hazards model are plotted (solid line) against the percentiles of the linear predictors from the same model. These predicted probabilities are termed the predicted risks, $R(v,t)$, where v is the proportional rank of the linear predictors and t is the selected time. The horizontal dashed line represents the Kaplan-Meier estimate for remaining event-free at the same time, t, with no covariates taken into account ($p_0(t)$). The grey shaded area between the two lines then represents the value of $TG(t)$, that is the increase in accuracy of the predicted probability at time t through use of the Cox proportional hazards model, with covariates, as compared to the reference estimated probability from the Kaplan-Meier curve.

The value of TG at time t is calculated as

$$TG(t) = \int_0^1 |R(v,t) - p_0(t)| dv,$$

where $R(v,t)$ denotes the predicted probability from the Cox proportional hazards model, $Pr[T > t|v]$, v is the proportional rank of the linear predictors, and $p_0(t)$ is the Kaplan-Meier estimate of remaining event-free at time t. Since the predicted probabilities are

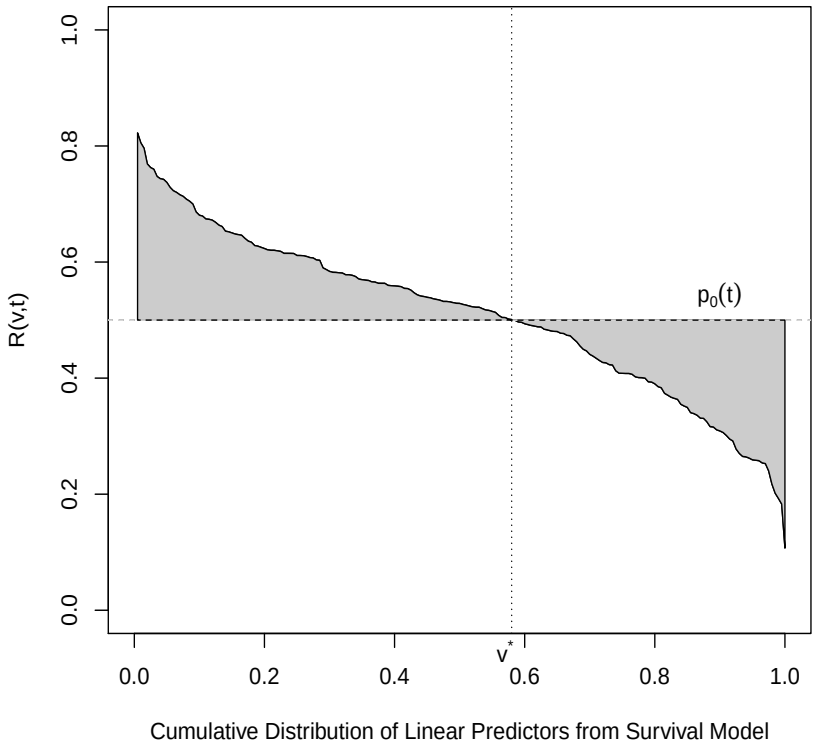

Figure 5.1: Hypothetical Example of $TG(t)$ with one continuous covariate

distinct for each value of the linear predictors and form a step function, the value of TG can be estimated by summing the individual differences between the two curves for each value of the scaled linear predictor, such that

$$TG(t) = \sum_{0}^{v^*}(R(v,t) - p_0(t)) + \sum_{v^*}^{1}(p_0(t) - R(v,t)),$$

where v^* represents the point at which the two lines intersect (seen in Figure 5.1).

For both the binary and survival definitions of TG, there is an upper bound which occurs when there is complete separation of outcomes (yes or no/survival or not) across the range of linear predictors, such that knowledge of the covariates and respective coefficients guarantees knowledge of the outcome. The estimate of TG must then be scaled for this maximum value, to achieve values within the range $[0, 1]$. For survival data, this must occur when the event status of all patients is known, when there is no censoring. This is illustrated in Figure 5.2, where the separation of outcomes is demonstrated through the predicted probability of one for a proportion $p_0(t)$ of the sample, and a value of zero for the remaining $(1 - p_0(t))$ of the sample. With no censoring, this value of $p_0(t)$ corresponds exactly to the average event-free probability from the Kaplan-Meier function as described above, since this proportion of patients remains without an observed event at time t.

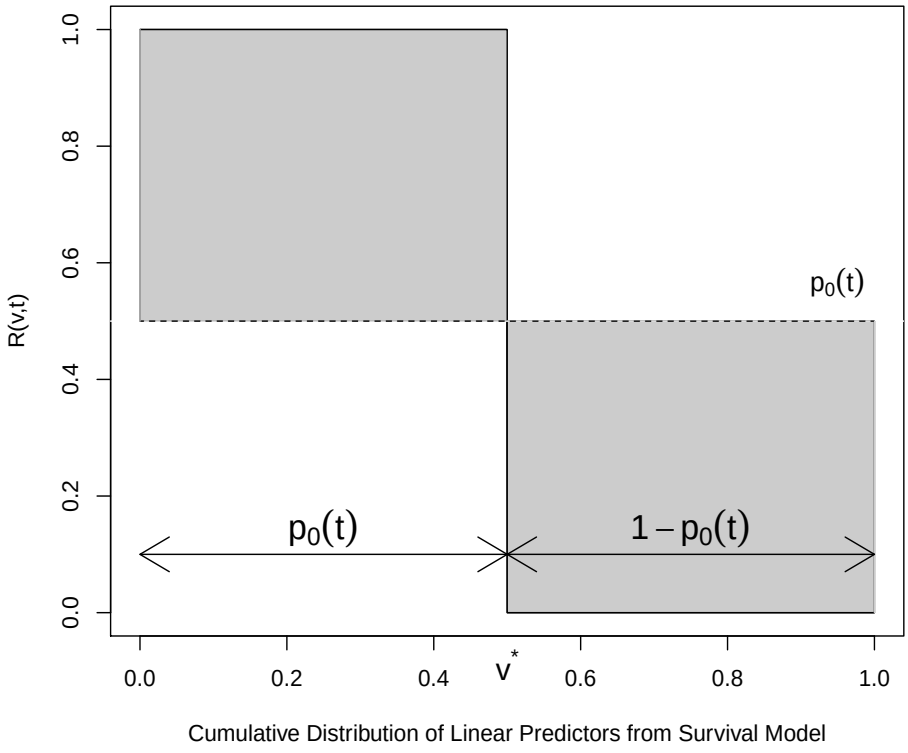

Figure 5.2: Maximum $TG(t)$

The maximum value of $TG(t)$ is then derived as the sum of the two grey-shaded areas in Figure 5.2, by

$$\begin{aligned} TG(t)_{max} &= \sum_{0}^{v^*}(R(v,t) - p_0(t)) + \sum_{v^*}^{1}(p_0(t) - R(v,t)) \\ &= (1 - p_0(t))p_0(t) + p_0(t)(1 - p_0(t)) \\ &= 2p_0(t)(1 - p_0(t)) \end{aligned} \tag{5.1}$$

and so a standardised, or scaled, version of $TG(t)$ is proposed for use, defined as

$$TG_{STD}(t) = \frac{TG(t)}{2p_0(t)(1 - p_0(t))}, \tag{5.2}$$

and Choodari-Oskooei et al. (2015) recommend bootstrap resampling to construct confidence intervals for the measure. Some of the benefits of this $TG_{STD}(t)$ measure include the ease of computation, particularly since the quantities needed can be calculated very simply from standard software packages. The final value of $TG_{STD}(t)$ lies between 0 and 1, allowing simple interpretation for non-statisticians as well as informal comparison to other association measures lying on this same scale. Further, when considering the use of the measure from the perspective of surrogate endpoint evaluation at the individual-level, it is possible to include time-dependent covariates that can reflect changing disease states (e.g. no progression to progression). In fact, the use of the Kaplan-Meier function and Cox proportional hazards model is consistent with estimation within the information theory approach to evaluating individual-level surrogacy described in Chapter 4, with only the final stages of the approaches differing in how the parameters estimated from these models are summarised.

Choodari-Oskooei et al. (2015) conducted simulation exercises to thoroughly assess the performance of $TG_{STD}(t)$ when used to build predictive models, including the impact of censoring, covariate distributions (including time-dependent covariates) and strength of covariate effects. Most notably, the measure is found to be independent of random censoring, which is a strong advantage when considering survival data. Further, the measure is found to increase with increasing strength of covariate effect, therefore reflecting the

increased association between an outcome and set of one or more covariates. In simulations, $TG(t)$ was found to plateau at a value of approximately 0.5, which led to values of $TG_{STD}(t)$ close to one for large covariate effects, suggesting that it is possible for the measure to reflect large associations between outcomes and covariates.

One highlighted feature of $TG_{STD}(t)$ is that it can be measured at a specific point in time. This allows for calculation and comparison of the measure across multiple studies with differing periods of observation, since the fixed time, t, can be selected such that it is relevant to all studies. This also introduces some element of subjectivity, since the value may not remain constant over the full period of observation. The simulation study of Choodari-Oskooei et al. (2015) demonstrated that the pattern of $TG_{STD}(t)$ over time can depend on the effect size and distribution of the covariate. To address this, sensitivity analyses should be conducted to evaluate $TG_{STD}(t)$ across a range of times. Additionally, specific times that are most relevant to the disease being studied can also be selected. Indeed, Choodari-Oskooei et al. (2015) highlight the need for careful selection of the timepoint of evaluation, and this topic will be further discussed in subsequent sections.

Based on the simulations, Choodari-Oskooei et al. (2015) recommend $TG_{STD}(t)$ as a measure of association, and with the benefits highlighted above it also appears worthy of further consideration as a potential approach for the evaluation of surrogate endpoints. The interpretation of $TG_{STD}(t)$ is highly relevant to a surrogacy setting, since it would be possible to use the measure to determine whether, and by how much, predictions of a long-term true clinical endpoint could be improved through knowledge of the surrogate endpoint as a covariate. The suitability of $TG_{STD}(t)$ as developed by Choodari-Oskooei et al. (2015) as an individual-level surrogacy evaluation measure is therefore discussed in Section 5.4, with continued focus on the evaluation of time-to-event surrogate endpoints of time-to-progression and progression-free survival for the true endpoint of overall survival. First, selection of an appropriate timepoint at which to measure $TG(t)_{STD}$ is discussed.

5.3.3 Selection of t

As discussed above, $TG(t)$ and $TG_{STD}(t)$ are time-dependent measures of association, with t selected as the timepoint at which to assess the relationship between covariates and outcome. There appears, therefore, to be an arbitrary choice of timepoint, with Choodari-Oskooei et al. (2015) noting that the value of $TG_{STD}(t)$ may increase over time, and specifically that there is minimal discrimination of predictions in event-free probabilities near the origin and near maximal timepoints. Choodari-Oskooei et al. (2015) recommend that the timepoint selected for evaluation of $TG(t)$ and $TG_{STD}(t)$ should be clinically relevant to the disease under study, and will therefore vary depending on the application of the methodology to each specific setting. Importantly, the ability to choose the timepoint allows for estimation of the association across studies with differing periods of follow-up, which is an advantage for survival studies, which may differ substantially in the maturity of data at the time of analysis.

When investigating diseases with specific timepoints of relevance, such as those where treatment has a curative intent and the number of OS events is expected to reduce significantly after a certain timepoint, there will generally be clinical consensus as to which timepoint is most useful. For other disease settings, it will need to be a discussion between statistician and clinical expert, to establish a point at which the data are mature and robust enough to make inferences, whilst remaining relevant. One summary statistic that is frequently used in oncology indications is median survival; the earliest time at which the probability of remaining alive drops below a value of 0.5. Given that this is a key parameter used and understood by statisticians and clinicians, further consideration of $TG_{STD}(t)$ in this thesis focuses on estimation at the time of median OS. Selection of the median OS time also allows for data relating to the surrogate endpoint to be reasonably mature, when assuming that the surrogate outcome will be reached sooner than that of the true endpoint. Additional sensitivity analyses of other percentiles of the OS distribution (20% to 80%) are also considered to assess the sensitivity of the measure to changes in

both the timepoint and the amount of information available (see Section 5.6). As noted, for diseases where there is a clinically relevant timepoint of interest, or indeed when the time to median survival is so long that it is not feasible to reach in a reasonable period of patient observation, alternative choices should be taken.

5.4 $TG_{STD}(t)$ as a Measure of Individual-Level Surrogacy

Choodari-Oskooei et al. (2015) proposed the $TG_{STD}(t)$ measure in the context of building and evaluating prognostic models, to identify factors that could be considered useful in predicting survival status at a given time. However, the predictive ability of a set of covariates in determining survival status is highly relevant to surrogate endpoint evaluation, where, at the individual level, it is of interest to identify whether an intermediate disease state can reliably predict true long-term outcome. A high predictive ability would indicate that knowledge of the surrogate outcome can allow for reliable prediction of the true endpoint, thereby allowing the surrogate outcome to be used for the purpose of regulatory decision making. A low predictive ability would suggest that the surrogate cannot reliably predict long-term outcome, and therefore should not be used as a primary endpoint in confirmatory clinical studies. Furthermore, the $TG_{STD}(t)$ method has been shown to be independent of random censoring, is increasing with increased association between a covariate and outcome, and lies within the range of zero to one, allowing indirect comparison with the surrogacy approaches investigated within this research and commonly used in practice.

In the context of individual-level surrogacy evaluation, the aim is to determine whether a surrogate outcome can reliably predict an unobserved long-term outcome, after accounting for treatment. Covariates of interest are therefore treatment, the time-dependent surrogate outcome, and any additional prognostic or patient characteristic factors thought to influence the true outcome. As for previous methods examined in this thesis, further prognostic factors are not considered and the focus remains on accounting for treatment only.

In the setting of interest here, namely assessment of a time-to-event surrogate for a time-to-event true endpoint, we focus on the survival representation of $TG(t)$, and consider the surrogate endpoint as a time-dependent covariate. For consistency with both Choodari-Oskooei et al. (2015) and estimation of the information theory measure described in Section 4, Cox proportional hazards models are used to estimate coefficients for the covariates of treatment and the time-dependent surrogate outcome, and the Kaplan-Meier function is used as an estimate of survival without covariates. Application of $TG_{STD}(t)$ in the new context of surrogate endpoint evaluation aims to provide insight into whether the method is able to adequately capture strengths of association between surrogate and true endpoints. Results that are within the range of, but ideally higher than, those based on the information theory method would lead to further consideration of $TG_{STD}(t)$ as an approach to evaluating the predictive ability of surrogate endpoints.

For this initial examination of $TG_{STD}(t)$ as a measure of individual-level surrogacy, the models are constructed in the same way as described for the information theory method (Section 4.2.3), and further re-capped here. In order to include the surrogate endpoint as a time-dependent covariate, to reflect the change in disease status (from no disease progression to disease progression), the time period from baseline to the true endpoint of overall survival, $[0, T)$, is separated into two intervals that reflect the potential change in surrogate outcome at time S; $[0, S)$ and $[S, T)$. During the first interval there is no progression, and during the second interval there may or may not be disease progression dependent on the progression status at time S. Time-dependent indicator variables are used to denote disease status both at time T and during the interval $[S, T)$. When based on a meta-analysis, $TG_{STD}(t)$ would be calculated for each trial individually, and combined across trials using a weighted average, using the sample size or number of events.

Consistent with previous chapters, both time-to-progression (TTP) and progression-free survival (PFS) are considered as potential surrogates for OS. When TTP is used as the surrogate, the status during the interval $[S, T)$ is based only upon the disease status provided by the surrogate endpoint at time S. Therefore, if a patient experiences disease

progression at time S, this is accounted for in parameter estimation through a change in the value of the time-dependent covariate from zero to one. If a patient does not experience disease progression during their period of observation, their time-dependent covariate remains at a value of zero across the entire interval $[0, T)$. For PFS, in cases where patients had death without prior progression, the interval $[S, T)$ was assumed to have length of one day, so that the data reflects the occurrence of the surrogate event.

For this setting, the Cox proportional hazards model used to predict the survival probability therefore contains two binary covariates; treatment and the time-dependent surrogate outcome of disease progression or not. The predicted survival probability based on these two binary covariates forms a step function, with four distinct levels representing the range of possible combinations for the linear predictor (see the example in Figure 5.3). Since the surrogate outcome is represented as a time-dependent covariate, the linear predictor used for the plot and in calculation of $TG(t)$ and $TG_{STD}(t)$ also depends on time. Whilst the covariate coefficient remains constant, the change in covariate value leads to a change in linear predictor value, and so the value used to calculate $TG(t)$ must reflect the covariate status at the chosen time, t. Based on this example figure, estimation of $TG(t)$ is taken as the sum of the areas of the four individual sections shown in grey. The maximum value of $TG(t)$ is calculated using Equation (5.1) and $TG_{STD}(t)$ can subsequently be calculated according to Equation (5.2). To examine the performance of $TG_{STD}(t)$ in evaluating individual-level surrogacy, a simulation study is used, described in the next section.

5.4.1 Description of the Simulation Study

A simulation study is required to assess the performance of $TG_{STD}(t)$ in the previously unexplored setting of surrogate endpoint evaluation. As for previous simulation studies presented in this thesis, it is important that the underlying strength of surrogacy can be adequately controlled. As for the information theory approach, it is very difficult to

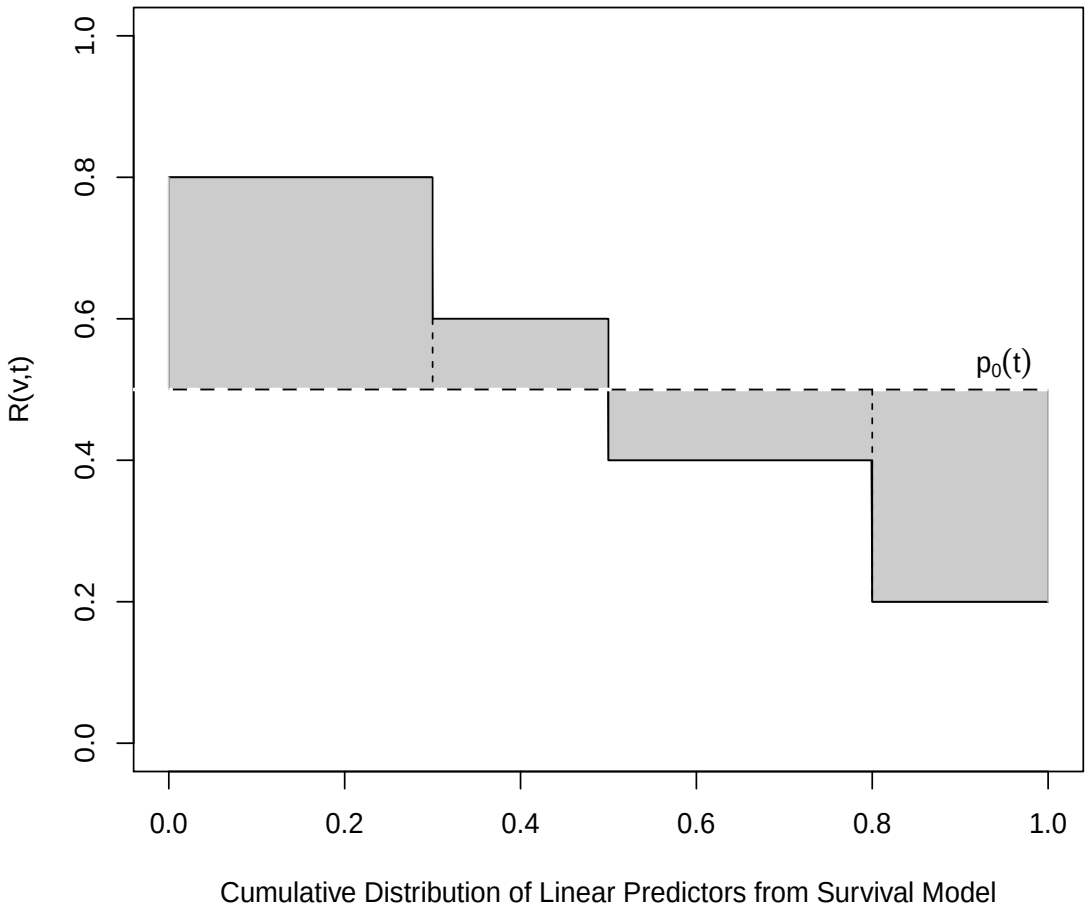

Figure 5.3: $TG(t)$ - example for two binary covariates

achieve this for this new setting, as each sample combines estimates from multiple modelling approaches. Therefore, to maintain consistency across all three simulation studies and to ensure comparability of results, the simulated datasets used in Chapters 3 and 4 for assessment of the previously examined surrogacy measures were used to explore the performance of $TG_{STD}(t)$. As for the information theory approach, whilst the value of τ may not perfectly reflect the true $TG_{STD}(t)$ measure of association, such an approach allows for overall control of the true individual (τ) and trial (R^2_{trial}) association levels, subject to sample variability, and conveniently allows indirect comparison with results of the previ-

ously examined surrogacy methods. It is likely that any practical application of surrogacy evaluation would be based on a number of different methods, including some sensitivity analyses, and a substantial limitation to interpretation of surrogacy would occur if the available methods gave conflicting results for the same dataset. As such, all three surrogacy measures examined in this thesis are based on identical simulated datasets, with 5,000 repetitions of each scenario of interest. Given the comparability of results for previous approaches, the range of simulation scenarios has been reduced slightly for investigation of $TG_{STD}(t)$, and the scenarios examined are displayed in Table 5.1. To combine the results across the number of trials, a weighted average of study specific estimates is used, weighted by trial size. Results of the simulation study are presented in the forthcoming section, first for TTP and followed by PFS.

Table 5.1: Simulation Scenarios

Factor	Scenarios under simulation
Surrogate Endpoint	TTP, PFS
Data Generation	Clayton, Gumbel
Number of trials	6
Number of patients per trial	80, 120
Trial-level association	0.5
Individual-level association	0.2, 0.5, 0.8
Censoring Rate (on T)	0%, 30%
Range of treatment effects*, σ	0.1

*Hazard ratios ranging $42\% - 203\%$ from the mean.

5.4.2 Results

Time to Progression

Results of the application of $TG_{STD}(t)$ to the datasets simulated according to Table 5.1 are presented in the form of boxplots, consistent with previous results presented in this thesis. Each combination of data generation algorithm (Clayton or Gumbel) and surrogate endpoint (TTP or PFS) is presented separately, with results for all sample sizes and censoring proportions included in each figure and differentiated with a legend (for sample size) or along the x-axis (censoring proportion). The strengths of individual-level surrogacy (0.2, 0.5 or 0.8) are presented as separate plots from left to right within each figure.

Dashed reference lines for the value of τ used for data generation are included for each scenario. As previously noted, since $TG_{STD}(t)$ is not estimating the value of association as expressed by the copula parameter τ, it is not expected that estimates will always be close to the input value. However, the values used in data generation are intended to demonstrate whether the approach can reliably identify poor from good surrogates. Whilst the absolute value of $TG_{STD}(t)$ may not fully match the intended τ, it is important to understand whether a reliable conclusion can be drawn, whether varied strengths of individual-level association can be differentiated by $TG_{STD}(t)$, and whether the method provides an estimate that is broadly comparable to the underlying association within the data.

Figure 5.4 contains estimates of $TG_{STD}(t)$ for the surrogate endpoint of TTP, calculated at the time of median OS and based on Clayton copula data generation. Results presented in this figure show a number of interesting features. Firstly, as concluded by Choodari-Oskooei et al. (2015), $TG_{STD}(t)$ appears to be unaffected by censoring, with minimal changes in the medians and ranges of estimates across the 5,000 simulation runs when approximately 30% of patients are censored, as compared to studies with no censoring. Secondly, the range of estimates for each simulation setting is reasonably small, with

very limited overlap between the differing strengths of individual-level association. In previous chapters, it was found for both the two-stage meta-analytic copula and information theory methods that the ranges of estimates often overlapped between the three levels of τ investigated, which would hamper interpretation of the results in practice. $TG_{STD}(t)$ does not appear to suffer from this limitation, even for the setting of small sample sizes examined here. Increasing the sample size for each study from 80 to 120 patients also led to a reduction in the range of estimates, suggesting that this could be further improved with more data availability. Finally, despite there being some under-estimation of medium and high levels of association, the estimates of $TG_{STD}(t)$ are higher than those based on the information theory approach, suggesting that the measure could offer potential as an alternative approach in assessing individual-level surrogacy.

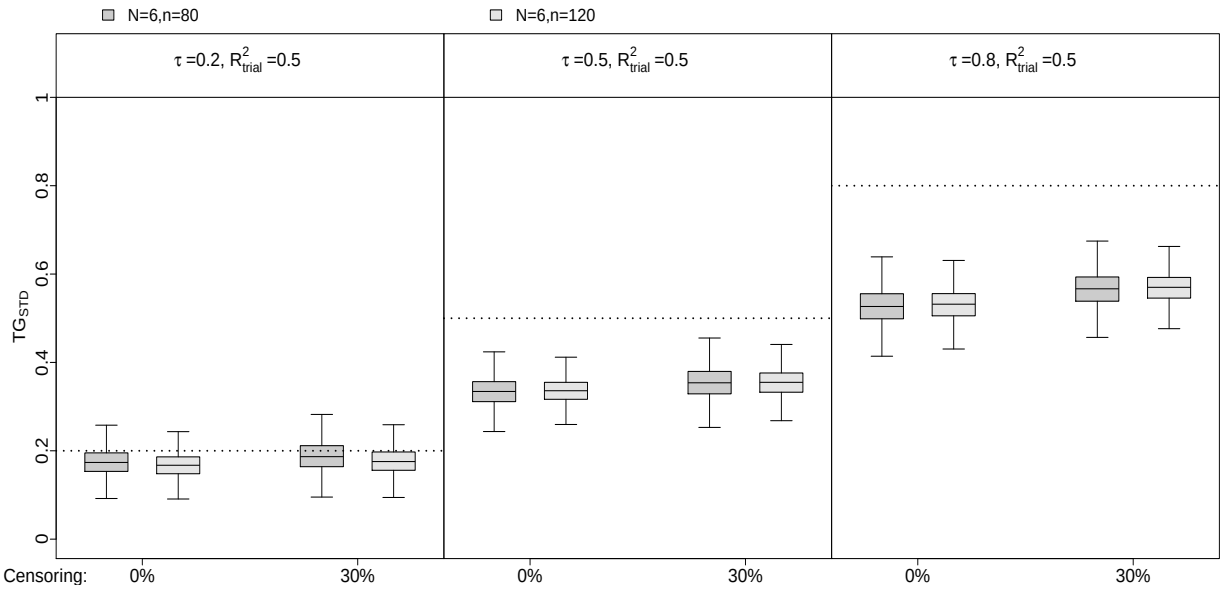

Figure 5.4: Boxplots of estimates of $TG_{STD}(t)$ at Median OS: TTP, Clayton Copula Data Generation, Total Gain Application

These findings were consistent when considering Gumbel copula generated data (Figure 5.5), which assumes a different dependence structure to that of the Clayton copula. As seen for the information theory method, the values of $TG_{STD}(t)$ are generally higher when based on the Gumbel copula generated data as compared to the Clayton copula,

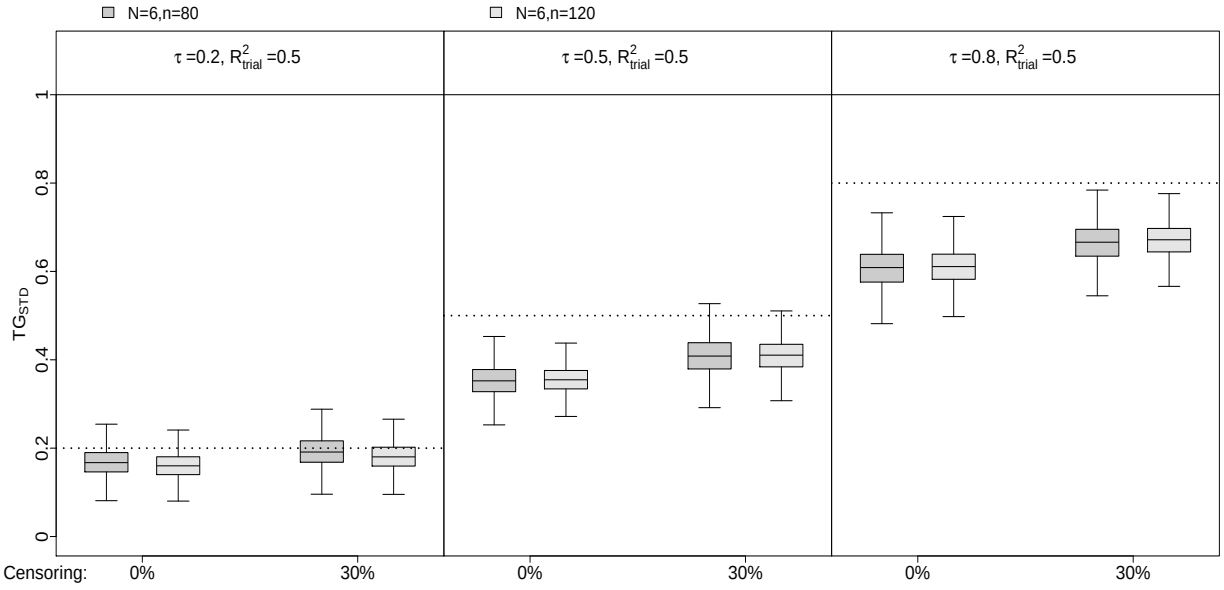

Figure 5.5: Boxplots of estimates of $TG_{STD}(t)$ at Median OS: TTP, Gumbel Copula Data Generation, Total Gain Application

reflecting the different dependence structure between the two endpoints assumed by these models. This reduces the level of under-estimation for medium and high levels of τ without leading to over-estimation of the lowest association value. Importantly, the highlighted advantages of $TG_{STD}(t)$ observed for the Clayton generated data also appear to be present for the Gumbel generated data, with limited overlap between estimates for the three investigated levels of τ, reasonably small ranges of estimates, and no notable impact from the introduction of censoring. Importantly, the estimates of $TG_{STD}(t)$ remain higher than those based on the information theory method, again suggesting that $TG_{STD}(t)$ is worthy of further consideration as an alternative measure of individual-level surrogacy.

Progression-Free Survival

Similar data presentations are provided in Figures 5.6 and 5.7 for Clayton and Gumbel generated data respectively, based on PFS. Each figure contains three plots; one for each value of τ used in the simulation, with the boxplots demonstrating the simulation results for both sample sizes and censoring proportions considered.

For both the Clayton and Gumbel data generation, estimates of $TG_{STD}(t)$ appear very promising, with median values being very near to the reference line for τ, with reasonably narrow ranges. Whilst τ is not to be considered the exact reference value, these results demonstrate that $TG_{STD}(t)$ is able to differentiate well between low, medium and high strengths of individual-level surrogacy, with no overlap of estimates for each of these values. Increasing the sample size from 80 to 120 patients also leads to slightly reduced ranges of estimates. This estimation performance is a key advantage of the $TG_{STD}(t)$ method, since all results presented for the two-stage meta-analytic copula and information theory methods, with the exception of TTP Clayton data for the copula method, demonstrated an overlap in ranges of surrogacy estimates that make it difficult to clearly differentiate between strengths of association, and in many cases would lead to erroneous conclusions.

As for the TTP setting, there is minimal impact when censoring is present in the data, with estimates increasing by only a negligible amount. Based on Clayton generated data, $TG_{STD}(t)$ slightly under-estimates the highest level of association, however this under-estimation is lower than for the TTP scenarios. As for previously explored scenarios, estimates based on Gumbel generated data are slightly higher than those from Clayton generated data, however in this setting the estimates of $TG_{STD}(t)$ do not appear to be over-estimating any of the reference levels of association between endpoints.

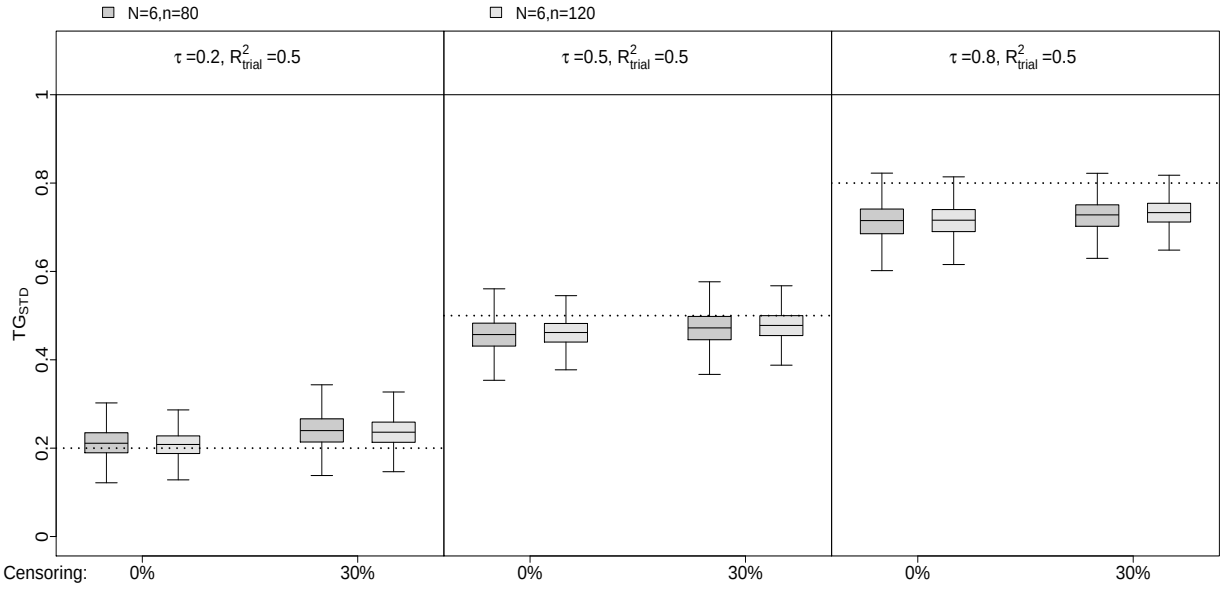

Figure 5.6: Boxplots of estimates of $TG_{STD}(t)$ at Median OS: PFS, Clayton Copula Data Generation, Total Gain Application

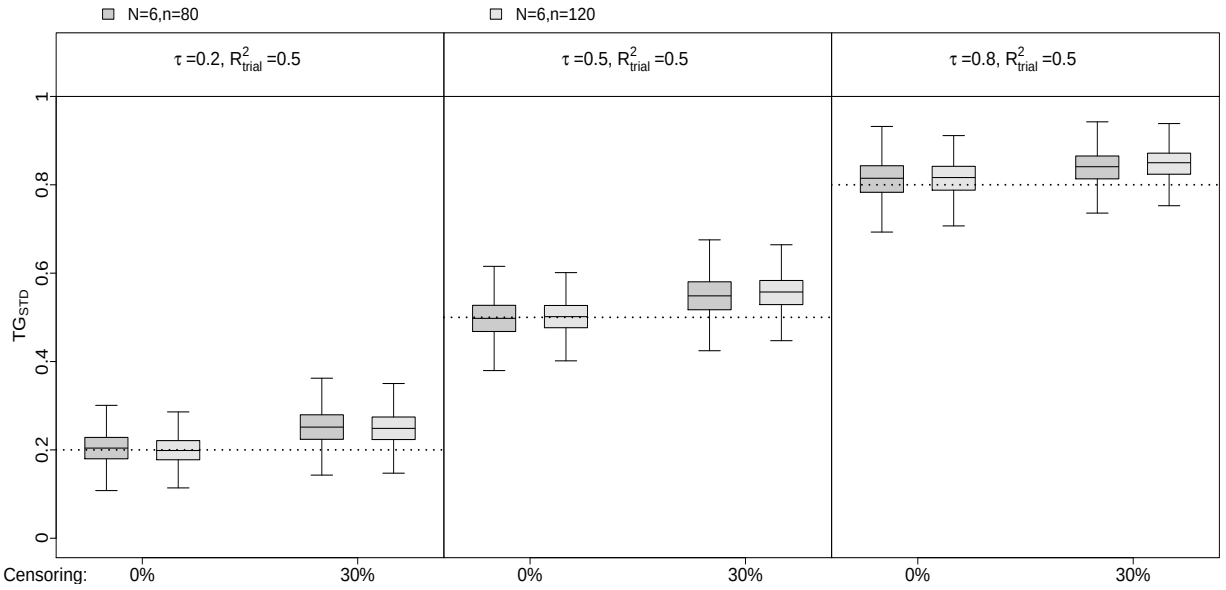

Figure 5.7: Boxplots of estimates of $TG_{STD}(t)$ at Median OS: PFS, Gumbel Copula Data Generation, Total Gain Application

The investigation of $TG_{STD}(t)$ in estimating the predictive ability of treatment and surrogate outcome has demonstrated that the measure has potential as a method for the evaluation of individual-level surrogacy, and results have demonstrated that $TG_{STD}(t)$

has many of the qualities that such a measure would require. For example, the measure appears to be independent of censoring, with estimates being very similar regardless of whether there are censored patients included in the data. Further, the measure increases with increased strength of association, as defined by τ, showing that increased association between a set of covariates and outcome can be reliably detected. In addition, the level of variability in the estimates appears to be reasonably low given the sample sizes tested, such that there is distinction between the various strengths of association. Given these benefits, $TG_{STD}(t)$ is considered worthy of further investigation as a measure of surrogate endpoint evaluation.

The examination of $TG_{STD}(t)$ conducted so far compares the predictive ability of a model containing both treatment and surrogate information to a null model which contains no covariates. However, in a surrogacy setting, it is of interest to understand the predictive ability of the surrogate endpoint after already accounting for any treatment effect on the surrogate and true endpoints. In order to use the concept of Total Gain for this purpose, further development of the approach is necessary. A key consideration in such development is whether a new measure could maintain the separation in the ranges of estimates across the different strengths of association, even for the setting of small sample sizes being investigated in this research. It is also of interest to determine whether there is any change on the impact of censoring, which remains of key importance for survival data. The next section discusses this further, and introduces a new version of $TG_{STD}(t)$ that can adequately adjust for the treatment effect on the true endpoint while maintaining the ease of calculation and conceptual appeal.

5.5 Extending $TG_{STD}(t)$ for Improved Surrogacy Evaluation

The version of $TG_{STD}(t)$ proposed by Choodari-Oskooei et al. (2015) compares a model with covariates to a null model, and so in the surrogacy setting a model containing both treatment information and surrogate information is compared to a model with no informa-

tion on either of these parameters. Since the true endpoint is also subject to a treatment effect, it is considered critical that any surrogacy assessment is able to quantify the additional predictive ability that comes from inclusion of the surrogate endpoint in a model which already contains treatment. The current form of $TG_{STD}(t)$ is unable to address this, and so may be over-estimating the predictive ability of the surrogate outcome. An extension of the approach is therefore necessary to maximise the potential for use in practice, and a proposal for such an extension is provided in this section.

In order to address the need to account for the treatment effect on the true endpoint, it is necessary to find a way in which the null model can be replaced with a model containing treatment information. This would allow for the predicted probability from the Cox proportional hazards model containing treatment and surrogate information to be compared to a model containing treatment, and would therefore provide quantification of the additional improvement in prediction that arises from knowledge of the surrogate outcome. The reference probability, denoted $p_0(t)$, would therefore need to be adjusted for treatment assignment, while the predicted probability from the Cox proportional hazards model would remain unchanged. Since the reference probability is based on the average probability of remaining event-free for all patients, through use of a Kaplan-Meier function, one possibility is to replace this with a Kaplan-Meier function stratified by treatment, which provides an average probability of remaining event-free within each treatment group separately. $TG_{STD}(t)$ could then be calculated as the difference between the predicted probability from the Cox proportional hazards model, and the respective Kaplan-Meier estimate depending on which treatment group the patient is assigned to. This provides an estimate of the ability of the surrogate (after accounting for treatment) to predict the true endpoint (after accounting for treatment). Such a change in methodology would provide a more accurate reference value for the impact of treatment alone on outcome, leading to a more informative quantification of whether a potential surrogate endpoint is truly predictive of the true endpoint. For this newly extended setting, $TG(t)$ is denoted $TG_Z(t)$, and $TG_{STD}(t)$ is denoted $TG_{STD,Z}(t)$, where Z denotes the treatment group.

This enhancement of the Total Gain methodology requires a change to the estimation approach, with the average, or reference, probability $p_0(t)$ being replaced with the value for each treatment group, $p_1(t)$ (for $Z = 1$) or $p_2(t)$ (for $Z = 2$). Extension to more than two treatment groups is also possible using the stratified Kaplan-Meier function, however notation will consider a binary treatment covariate for simplicity. As shown in Figure 5.8, the summation in Equation (5.3.2) would then be replaced with a summation across the separate treatment groups:

$$\begin{aligned} TG_Z(t) &= \sum_{0}^{v_1}(R(v,t) - p_2(t)) + \sum_{v_1}^{v_2}(p_2(t) - R(v,t)) \\ &+ \sum_{v_2}^{v_3}(R(v,t) - p_1(t)) + \sum_{v_3}^{1}(p_1(t) - R(v,t)), \end{aligned}$$

where v_1, v_2, and v_3 are the percentiles of the linear predictors that correspond with an intersection of the two probabilities and therefore a need to change the sign when calculating the difference between the two probabilities. Figure 5.8 is the same as the hypothetical example provided for $TG(t)$, however the Cox proportional hazards model is now assumed to contain treatment as well as one continuous covariate, and the dashed reference line is separated into two values to represent the stratified Kaplan-Meier estimates; one for each treatment group ($p_1(t)$ and $p_2(t)$). In this example, $p_1(t)$ and $p_2(t)$ are presented as being distinct across the distribution of linear predictors, such that treatment group is the strongest predictor and drives the ordering of the x-axis values. Should this not be the case, with other covariates causing larger differences to the linear predictors, the graphical display would look different, but the method for calculation of $TG_Z(t)$ would remain the same.

This additional development requires derivation of a new maximum value for $TG_Z(t)$, since perfect prediction of event status must now be considered separately for the two treatment groups, which have potentially different average probabilities of remaining event-free. As for the original $TG(t)$ measure, perfect prediction occurs when there is complete separation of outcomes (e.g. survival or not) across the range of linear predictors, with all

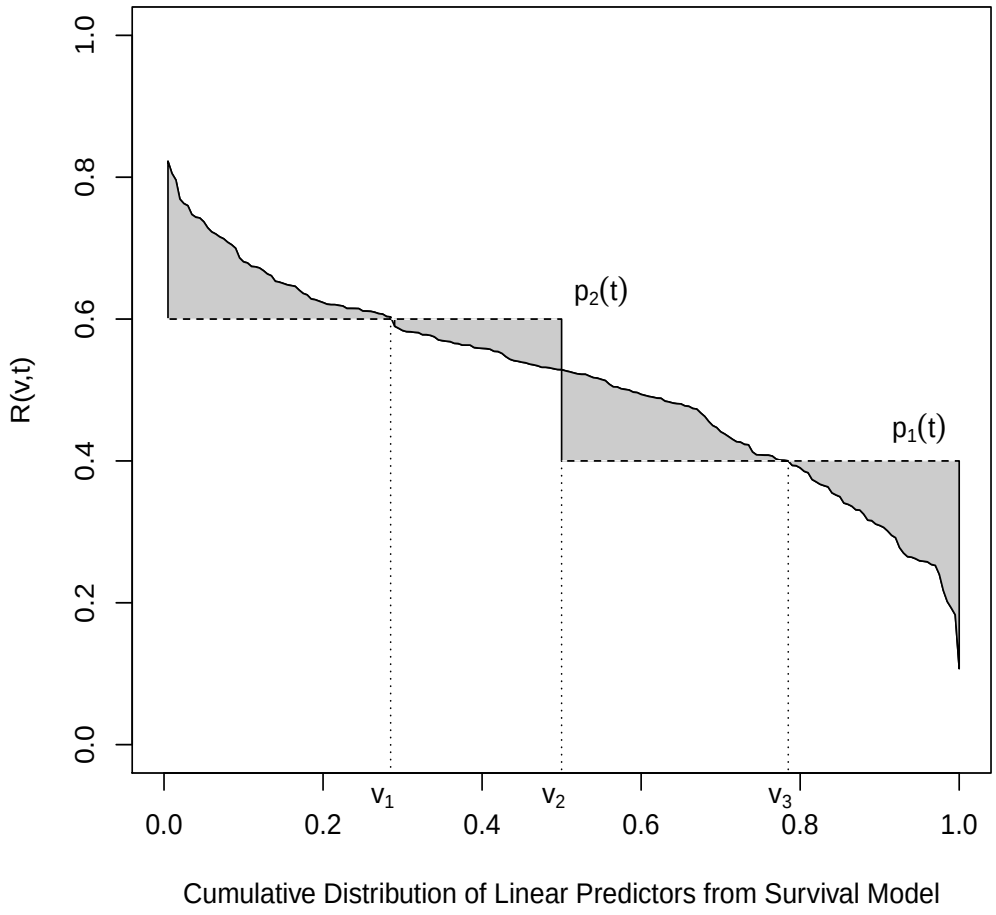

Figure 5.8: Hypothetical Example of $TG_Z(t)$ with treatment plus one continuous covariate

those remaining event-free in each treatment group having predicted event-free probability of one, and all remaining patients having predicted event-free probability of zero, and no censored patients. The average probability of remaining event-free in each treatment group is then equal to the proportion of patients remaining event-free in that treatment group.

The maximum value of $TG_Z(t)$ is illustrated in Figure 5.9 as the grey shaded area. This area can be separated into four individual segments which represent, from left to right, the patients in treatment group two who have not yet experienced the event of interest

(e.g. who remain alive), those in treatment group one who have not yet experienced the event of interest, those in treatment group two who have experienced the event (e.g. who have died) and those in treatment group one who have experienced the event. The values on the y-axis are one for the first two groups (patients without observed events) and zero for the second two groups (patients with events), with the average probabilities, $p_1(t)$ and $p_2(t)$, illustrated by the dashed lines. The width of each of the four segments corresponds to the proportions of patients in each treatment group who have experienced the event of interest or not at time t.

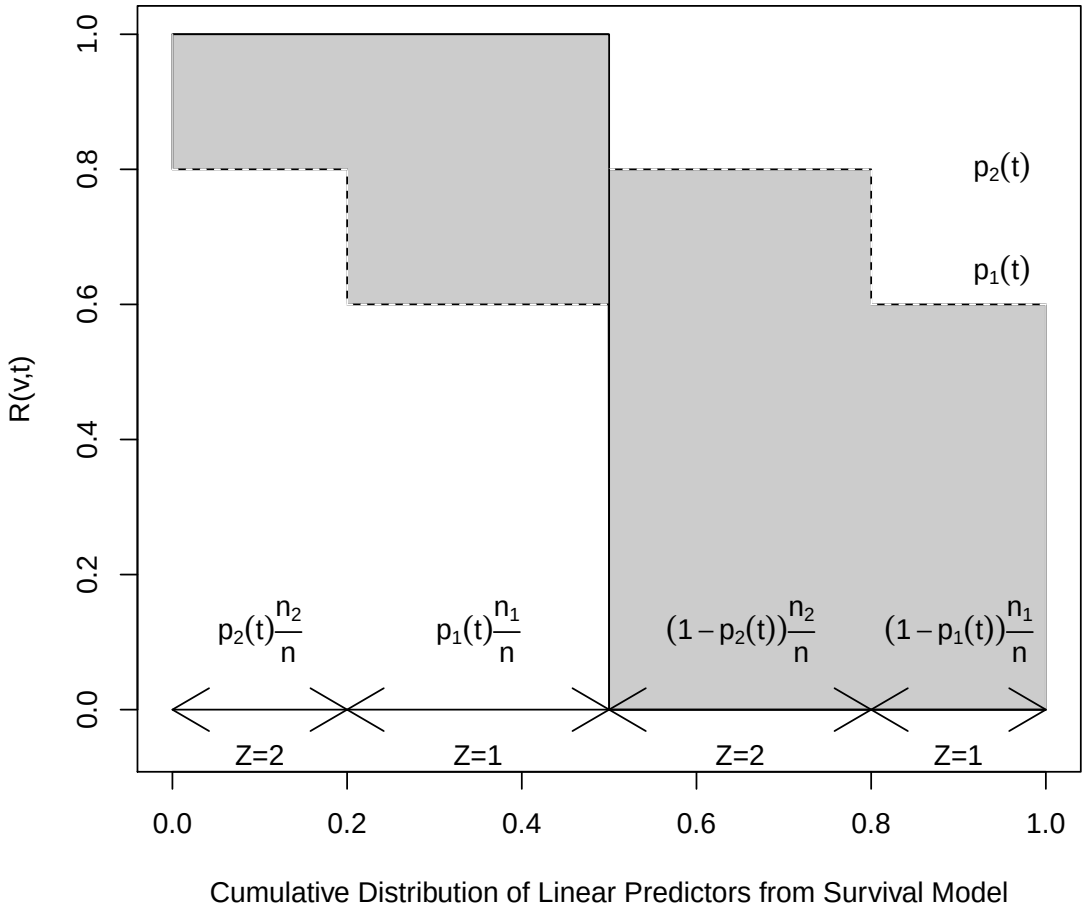

Figure 5.9: Maximum $TG_{STD,Z}(t)$

Suppose that the number of patients in treatment group one ($Z = 1$; control treatment) is n_1, and the number of patients in group two ($Z = 2$; experimental treatment) is n_2, such that $n_1 + n_2 = n$. Then, the proportion of patients in each treatment group who remain event-free at time t is equal to the probability of being event-free at time t for that group, multiplied by the proportion of patients being assigned to that group, such that:

$$\begin{aligned}
\text{Event-free in group two} &= p_2(t)\left(\frac{n_2}{n}\right), \\
\text{Event-free in group one} &= p_1(t)\left(\frac{n_1}{n}\right), \\
\text{Experienced the event in group two} &= (1-p_2(t))\left(\frac{n_2}{n}\right), \\
\text{Experienced the event in group one} &= (1-p_1(t))\left(\frac{n_1}{n}\right).
\end{aligned}$$

as illustrated in the graphic. The maximum value of $TG_Z(t)$ can then be calculated as the sum of the individual areas of each of these four grey sections, as:

$$\begin{aligned}
TG_Z(t)_{max} &= p_2(t)\left(\frac{n_2}{n}\right)(1-p_2(t)) + p_1(t)\left(\frac{n_1}{n}\right)(1-p_1(t)) \\
&\quad +(1-p_2(t))\left(\frac{n_2}{n}\right)(p_2(t)) + (1-p_1(t))\left(\frac{n_1}{n}\right)p_1(t) \\
&= 2p_2(t)(1-p_2(t))\left(\frac{n_2}{n}\right) + 2p_1(t)(1-p_1(t))\left(\frac{n_1}{n}\right).
\end{aligned}$$

Calculation of a standardised version of $TG_Z(t)$ is then possible using

$$TG_{STD,Z}(t) = \frac{TG_Z(t)}{2p_2(t)(1-p_2(t))\left(\frac{n_2}{n}\right) + 2p_1(t)(1-p_1(t))\left(\frac{n_1}{n}\right)}.$$

As for the original measure, confidence intervals can be constructed using bootstrap resampling. This development of the Total Gain concept allows the measure to be used in a surrogacy setting, by providing a method in which the improvement in prediction of the true outcome can be quantified through knowledge of the surrogate outcome. A value of $TG_{STD,Z}(t)$ close to zero would indicate that knowledge of the surrogate endpoint offers no additional accuracy in prediction of the long-term patient outcome, and suggests that the long-term outcome should remain as the primary endpoint for future clinical studies. A value of $TG_{STD,Z}(t)$ close to one, however, would suggest that prediction of

the true outcome is vastly improved through knowledge of the surrogate outcome, and would lead to high confidence that a surrogate endpoint could be used in practice. The simplicity of the original Total Gain concept remains with this new development, and importantly the measure remains easy to implement using standard statistical software. In order to establish whether the new modelling approach performs well in practice, an investigation has been undertaken using a simulation study, and this is introduced in the next section. The aims of this study are to identify whether $TG_{STD,Z}(t)$ can provide reliable estimates of individual-level surrogacy, particularly for the small sample sizes of interest in this research, to determine whether there are differences based on the choice of surrogate endpoint or underlying data structure, and to assess whether the measure remains robust to censoring.

5.5.1 Description of the Simulation Study

To thoroughly examine the proposed development in the estimation of $TG_{STD,Z}(t)$, further simulations were conducted for the same scenarios discussed in Section 5.4.1, shown again in Table 5.2. This range of scenarios allows for assessment of the performance of the new measure for different surrogate endpoints, data generation algorithms and dependence structures, strengths of individual-level surrogacy and under different proportions of censoring. A total of 5,000 repetitions of each scenario were conducted, and datasets generated are identical to those used for the assessment of all previously described methods.

This wide range of simulations allows a rigorous assessment of the performance of the new approach, and results are presented in the next section. Furthermore, a number of sensitivity analyses were also considered to assess the performance of $TG_{STD,Z}(t)$ in assessing surrogacy in a wider context, with consideration of larger sample sizes, larger treatment effects, alternative timepoints of estimation and alternative data generation mechanisms; these additional results are described and discussed in Section 5.6.

Table 5.2: Simulation Scenarios

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

5.5.2 Results

As for previous sections, the results of the application of $TG_{STD,Z}(t)$ to simulated datasets are presented in the form of boxplots, with one figure per combination of data generation method and endpoint. Within these figures, individual plots show results for each strength of individual association, for all sample sizes and censoring proportions. Dashed reference lines for the true value of τ used for data generation are included for each scenario. Results are presented for $TG_{STD,Z}(t)$ at the time of median OS, first for TTP and subsequently for PFS.

Time to Progression

Results for the setting of TTP based on Clayton copula generated data are provided in Figure 5.10. With the development of $TG_{STD,Z}(t)$, it could be expected that accounting for treatment in the Kaplan-Meier survival estimates ($p_1(t)$ and $p_2(t)$) would lead to estimates of Total Gain that are lower than those based on the $TG_{STD}(t)$. It would be reasonable to expect that the reduction in Total Gain would be the largest for the lowest strength individual-level surrogacy, since the surrogate is expected to have poorer predictive ability in this scenario as compared to when the association between surrogate and true endpoints

is very strong. Any reduction in values of $TG_{STD,Z}(t)$ as compared to $TG_{STD}(t)$ would therefore be expected to decrease with increasing τ.

The results in Figure 5.10 demonstrate that values of $TG_{STD,Z}(t)$ are reduced only very slightly when adjusting for the treatment effect on OS, with the largest decrease occurring for the lowest value of τ, as expected. However, even in this case, the reduction appears to be small, with values remaining across a similar range. Values of $TG_{STD,Z}(t)$ continue to be estimated with reasonable consistency, demonstrated by the range of estimates being similar to those from the $TG_{STD}(t)$ measure. Encouragingly, these ranges continue to have very little overlap between the various values of τ used for data simulation, suggesting that whilst there is some under-estimation of the medium and high levels of association, it is possible to distinguish between the three levels. This is a promising feature, as it allows for reliable conclusions regarding the usefulness of a potential surrogate endpoint. It potentially also helps to alleviate concerns of under-estimation, as the separation in ranges could be used to justify a minimum threshold above which a surrogate could be considered worthy of further investigation. Increasing sample sizes from 80 patients per trial to 120 per trial also led to the range of estimates of $TG_{STD,Z}(t)$ becoming narrower, suggesting that the method could perform even better when there are more data available.

When considering further the Gumbel copula generated data, the results follow a similar pattern. Estimates of $TG_{STD,Z}(t)$ are slightly lower than those for $TG_{STD}(t)$ when the true individual-level surrogacy is low, but there is negligible impact on the medium to high levels of association. As mentioned above, this is not unexpected, since the association between surrogate and true endpoints for $\tau = 0.5$ or 0.8 becomes strong, and likely overwhelms any effect of treatment on the true endpoint. For these higher levels of association, a slight increase in estimates under censoring leads to values slightly higher than $TG_{STD}(t)$, however this does not hamper interpretation, particularly with the ranges of results remaining fairly distinct across the three levels of association. As for all previous scenarios, the small increases in values based on Gumbel generated data as compared to Clayton generated data lead to a reduction in the under-estimation of the

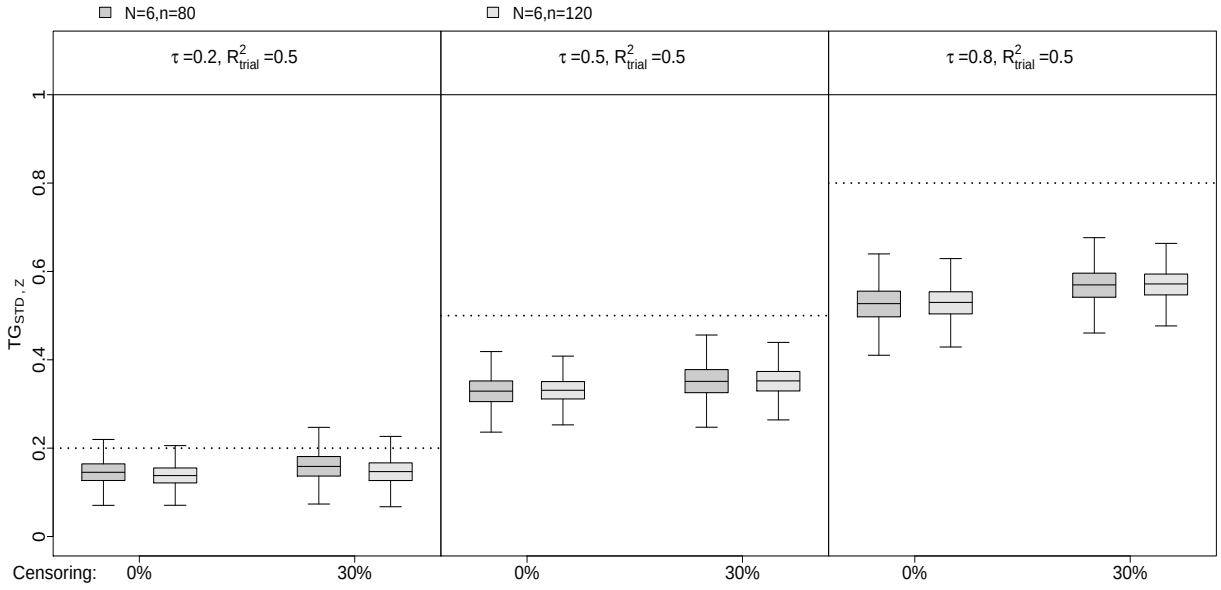

Figure 5.10: Boxplots of estimates of $TG_{STD,Z}(t)$ at Median OS: TTP, Clayton Copula Data Generation, Total Gain Application

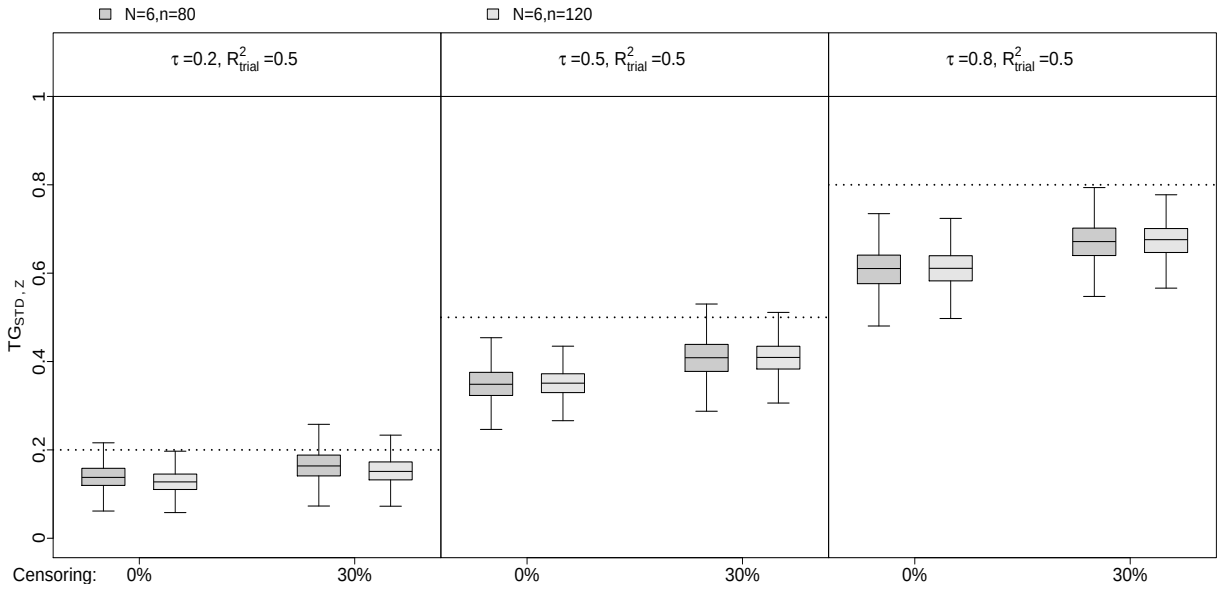

Figure 5.11: Boxplots of estimates of $TG_{STD,Z}(t)$ at Median OS: TTP, Gumbel Copula Data Generation, Total Gain Application

higher individual-level surrogacy values, without leading to over-estimation of the lowest individual-level surrogacy. Increasing sample sizes led the range of results to reduce slightly, suggesting that further improvement could be possible if more data were available.

Progression-Free Survival

Results for the Clayton copula generated PFS data are presented in Figure 5.12. As is expected, the estimates of $TG_{STD,Z}(t)$ are slightly higher than those based on TTP, and this leads to median values of $TG_{STD,Z}(t)$ that are reasonably close to the input value of τ. Ranges of estimates remain relatively small, and interestingly there is now no overlap in the results between the chosen values of τ, even for the smallest sample sizes. There appears to be minimal difference in results between those based on $TG_{STD}(t)$ and those of $TG_{STD,Z}(t)$, suggesting that accounting for the treatment effect on the true endpoint has little impact on the results or conclusions. This suggests that the predictive ability of the surrogate endpoint is not notably impacted by the assumed treatment effect; further discussion of this finding is provided in Section 5.6.3. As for the TTP scenarios, the estimates of $TG_{STD,Z}(t)$ appear to be marginally higher when there is censoring present in the data, but this has negligible impact on conclusions, and ranges of estimates reduce slightly with increased sample size.

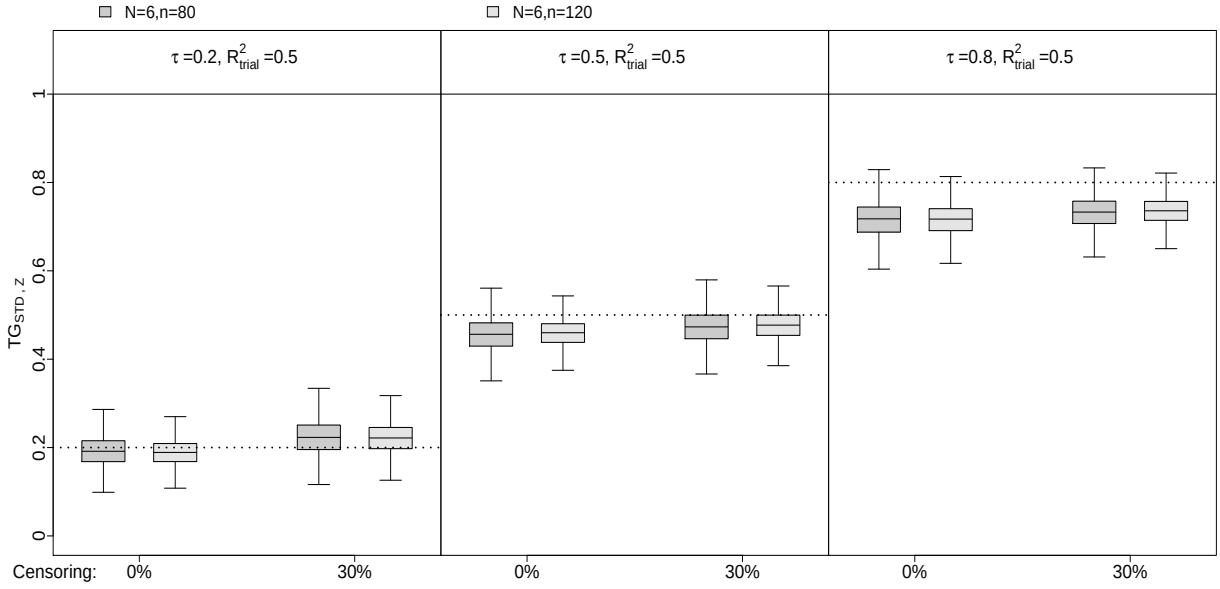

Figure 5.12: Boxplots of estimates of $TG_{STD,Z}(t)$ at Median OS: PFS, Clayton Copula Data Generation, Total Gain Application

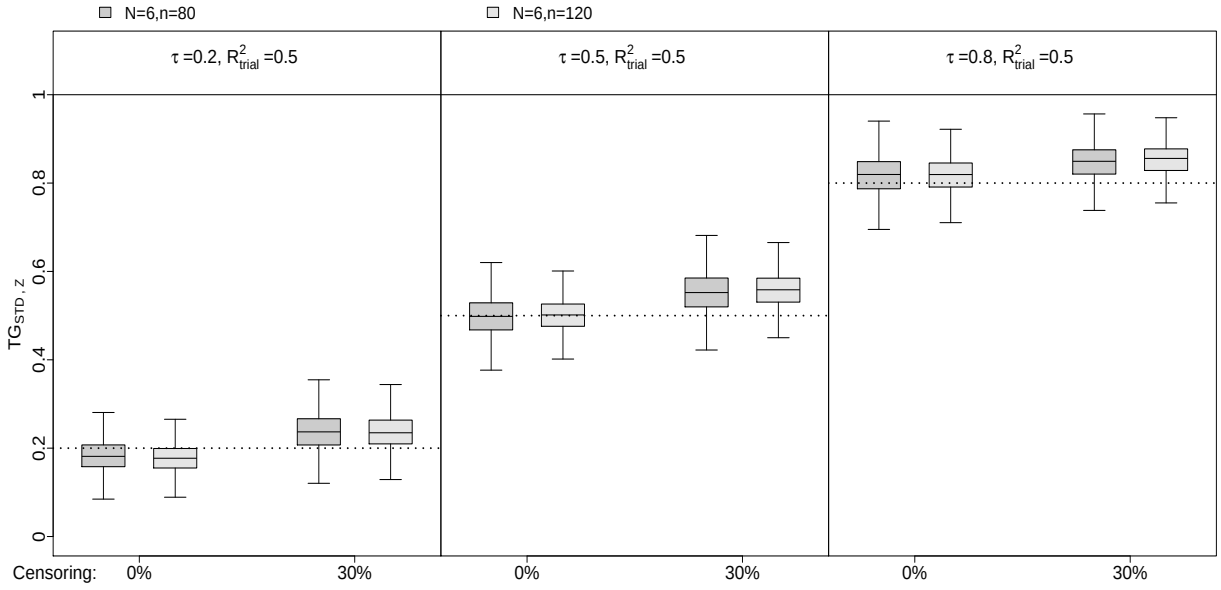

Figure 5.13: Boxplots of estimates of $TG_{STD,Z}(t)$ at Median OS: PFS, Gumbel Copula Data Generation, Total Gain Application

Further investigation based on Gumbel copula generated datasets (Figure 5.13) shows consistency with these findings, with results following the patterns previously observed; slightly higher estimates for Gumbel as compared to Clayton and slightly higher estimates for PFS as compared to TTP, but separation of the ranges of results across the three assumed levels of surrogacy. The increase in values of $TG_{STD,Z}(t)$ observed under censoring is slightly higher for the Gumbel generated PFS data than other scenarios, but again this does not hamper interpretation of results. Results based on datasets without censoring are very close to the reference value of τ, with those based on the censored datasets being very slightly higher. Again, there is very little difference between estimates of $TG_{STD,Z}(t)$ and $TG_{STD}(t)$, with differences being visible only for the lowest level of association. When increasing sample sizes, there is a slight reduction in the ranges of estimates, reflecting better estimation of $TG_{STD,Z}(t)$ with more data availability. Overall, the results for the assessment of PFS are very encouraging, and demonstrate that $TG_{STD,Z}(t)$ is able to clearly distinguish poor, mediocre and good surrogates, regardless of censoring and regardless of the underlying data structure.

Summary of Results

Results of the exploration of $TG_{STD}(t)$ and $TG_{STD,Z}(t)$ lead to the following conclusions:

- Total Gain provides a measure of predictive ability that is easily interpreted and simple to calculate.
- $TG_{STD}(t)$ provides good estimation of individual-level surrogacy based on Kendall's τ, despite some under-estimation for TTP, but requires modification to reflect the predictive ability of a potential surrogate endpoint.
- The newly developed measure, $TG_{STD,Z}(t)$ continues to perform well, with some under-estimation of association between TTP and OS, but strong performance when considering PFS as the surrogate endpoint.
- Whilst there is no true reference value against which the estimates of $TG_{STD,Z}(t)$ can be compared, the values were very similar to the input value of τ when considering PFS.
- Ranges of estimates of $TG_{STD,Z}(t)$ were not overlapping for any of the investigated values of association, reliably differentiating the strength of individual-level surrogacy.
- $TG_{STD,Z}(t)$ appears largely unaffected by censoring, with minimal increases in values that do not hamper interpretation.
- Overall, $TG_{STD,Z}(t)$ has performed well with regards to individual-level surrogacy evaluation.

5.6 Sensitivity Analyses

Following these simulations, additional steps have been taken to further examine the new approach to investigate whether it is suitable for wider use. A number of sensitivity analyses were conducted to investigate the performance of $TG_{STD,Z}(t)$ under various alternative assumptions. Firstly, to assess whether use of the copula-generated datasets could lead to bias in estimation of $TG_{STD,Z}(t)$, further datasets were generated according to the lognormal distribution, as described in Section 4.2.2. For these additional investigations, all three levels of individual surrogacy were considered, for 6 trials each containing 120 patients, and for both no censoring and 30% censoring. Again, identical datasets were used across the surrogacy evaluation methods to ensure consistency.

Additional analyses were also conducted to assess the sensitivity of the method to the timepoint selected for analysis, through estimation at various percentiles of the Kaplan-Meier distribution for OS. Whilst the main simulation results are based on estimation of $TG_{STD,Z}(t)$ at the time of median OS, further analyses at percentiles from 20% to 80% are conducted to determine whether there is a relationship between the association measure and time/data maturity. These additional sensitivity analyses are conducted for the largest sample sizes, with 6 trials each containing 120 patients, under 0% and 30% censoring, for TTP and PFS and for both Clayton and Gumbel generated datasets.

Further, the sensitivity of $TG_{STD,Z}(t)$ to the strength of treatment effect in the trials is investigated, through increasing the treatment effects on both the surrogate and true endpoints. In the main simulation study, the hazard ratios for TTP/PFS and OS are maintained as ≈ 0.67 and ≈ 0.82 respectively. The additional analyses increase the magnitude of treatment benefit, to determine whether such an increase has any impact on the estimation of the predictive ability of the proposed surrogate endpoint. In these new scenarios, hazard ratios of ≈ 0.50 for TTP/PFS and ≈ 0.6 for OS are considered, and the simulations are conducted for all three levels of individual surrogacy (0.2, 0.5, 0.8), for no censoring and for 30% censoring, based on 6 trials each containing 120 patients, and for

both the Clayton and Gumbel data generation algorithms.

Finally, the new method is assessed under the ideal situation where there exist substantial data on which to base an assessment of surrogacy, with sample sizes increased to 10 trials each containing 500 patients. These are considered for both no censoring and 30% censoring, to assess whether this has any impact on estimation. Due to the strong consistency of results between the various data generation methods, these analyses were restricted to the Clayton generated datasets. Results of each of these sensitivity analyses are described next.

5.6.1 Lognormal Data

In order to determine whether the observed results are biased through the use of a copula model to generate clinical trial datasets, further simulations were conducted using a lognormal distribution, without the use of a copula, and are presented in Figure 5.14 with TTP on the top row and PFS on the bottom row. For ease of comparison, all three data generation methods are included in this figure, which includes results for $N = 6$ trials of $n = 120$ patients, under 0% and 30% censoring. The light blue boxes contain results of $TG_{STD,Z}(t)$ for the Clayton generated data, and the dark blue boxes contain results for the Gumbel generated data.

Based on TTP, shown in the top row, results from the three data generation methods are highly consistent when the true individual-level surrogacy is low to medium ($\tau =$ 0.2, 0.5), with no discernible difference in values of $TG_{STD,Z}(t)$ for any of the scenarios examined. As the true individual-level surrogacy increases to the highest level, the values of $TG_{STD,Z}(t)$ based on the Clayton generated data are slightly lower than those from the other two data generation algorithms. However, the overall conclusion from this slight variation in estimates is expected to be the same, and interpretation of the results is not considered to be hampered by this variability. Similarly for PFS, shown in the bottom row, estimates of $TG_{STD,Z}(t)$ are reasonably similar across the three data generation

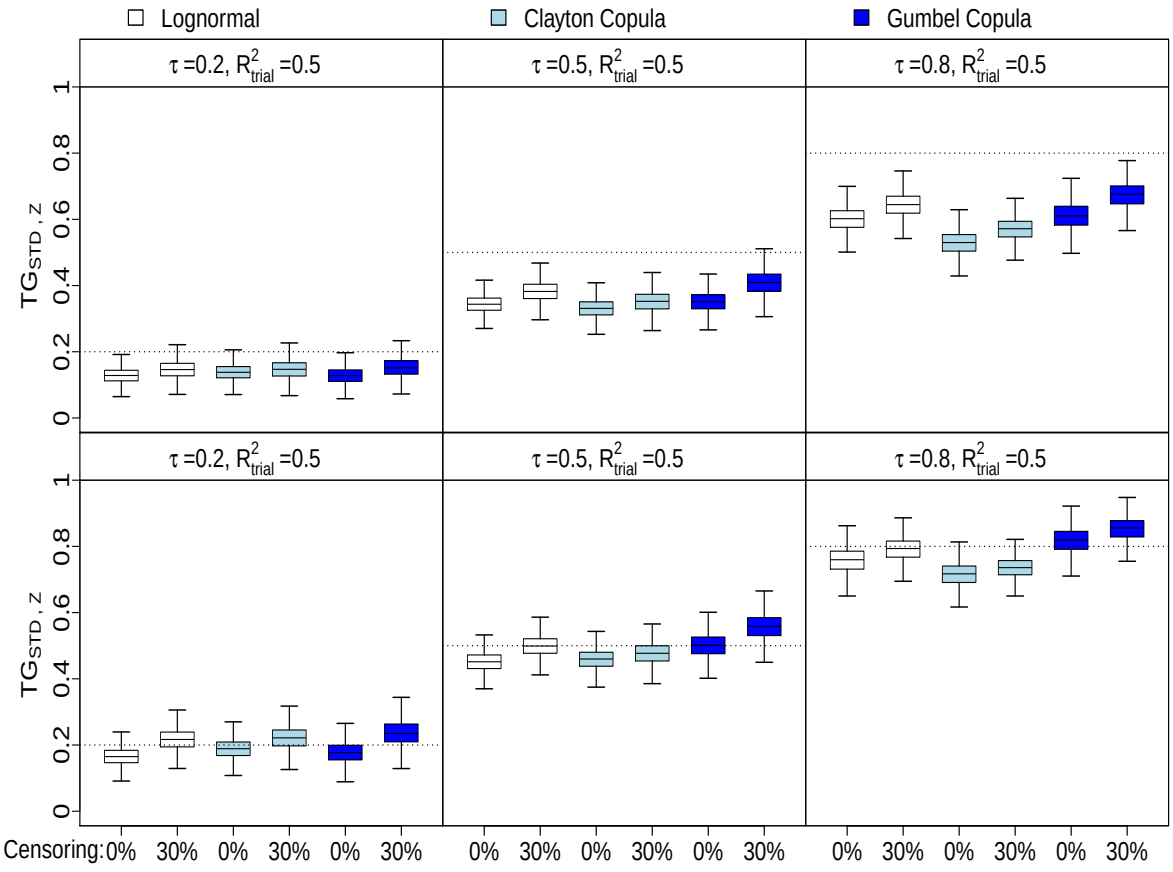

Figure 5.14: Boxplots of estimates of $TG_{STD,Z}(t)$ at Median OS: All Data Generation Methods, Total Gain Application, TTP (top row) and PFS (bottom row)

methods for all scenarios. For the lowest individual-level surrogacy, it is almost impossible to distinguish any differences in the estimates for datasets containing no censoring and datasets containing 30% censoring. As the strength of relationship between surrogate and true endpoints increases, values of $TG_{STD,Z}(t)$ based on Gumbel generated data appear to increase, with those based on lognormal data also increasing to a similar level when the individual surrogacy reaches the highest value of 0.8. The ranges of estimates remain mostly distinct across true underlying levels of τ, with values overlapping only for a very limited number of scenarios.

The overall consistency in results across these data generation algorithms therefore shows that use of the copula generated datasets in the simulation study has not favourably

or unfavourably biased the assessment of $TG_{STD,Z}(t)$ as a measure of individual-level surrogacy.

5.6.2 How does $TG_{STD,Z}(t)$ vary over time?

The $TG_{STD,Z}(t)$ measure is calculated at a fixed time that is selected based on relevance to the disease under study. In all simulations conducted so far in this thesis, t is chosen as the time of median overall survival for each individual study. To assess the impact of this, further calculation of the same simulated datasets was conducted using alternative percentiles of the overall survival Kaplan-Meier distribution, ranging from the early tail where 80% of patients remain alive, to the later tail where only 20% of patients remain alive. Due to similarity in the majority of results across the two different surrogate endpoints and two different data generation algorithms, only a selection of results is described in detail in this section. Full results for all scenarios examined can be found in Appendix Figures C.1 to C.4.

The scenarios selected for further discussion are those based on the highest strength of individual-level surrogacy ($\tau = 0.8$), since results appeared different depending on the combination of surrogate endpoint and data generation method. Estimates of $TG_{STD,Z}(t)$ for TTP and PFS are presented in Figures 5.15 and 5.16 respectively, with Clayton copula data presented on the top row and Gumbel copula generated data on the bottom row of each figure. In these figures, the Kaplan-Meier estimates along the x-axis reflect the proportion of patients who remain alive, and so decreasing values from left (80% of patients remain alive) to right (20% of patients remain alive) indicate increasing time of follow-up.

The pattern observed in the TTP data shown in Figure 5.15 is representative of the majority of scenarios examined, with values of $TG_{STD,Z}(t)$ increasing over time, similar to the pattern noted by Choodari-Oskooei et al. (2015). The variability in estimates also appears to increase very slightly in the upper (right) tail of the Kaplan-Meier distribution. This pattern is consistent between datasets with no censoring and those with 30%

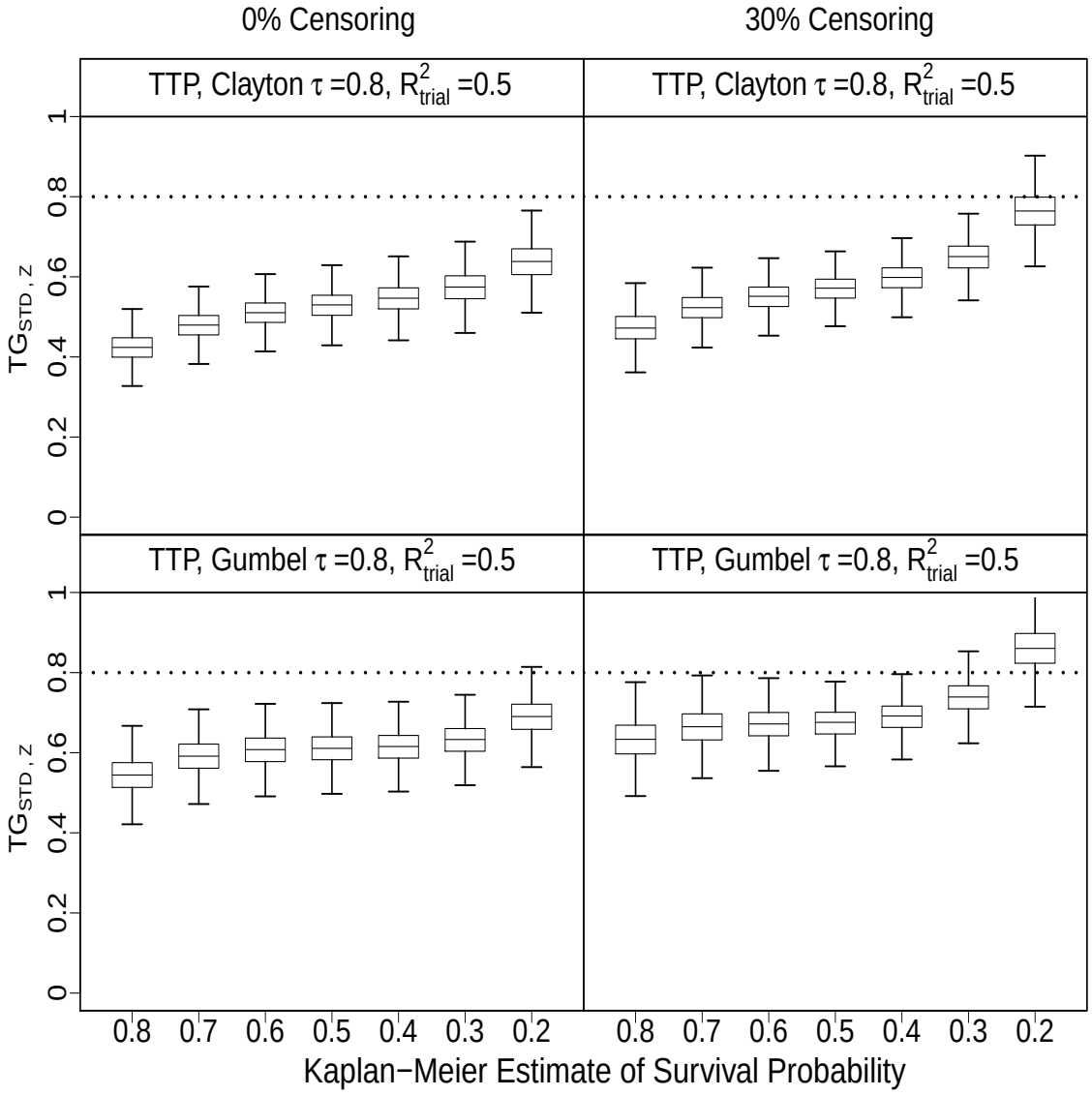

Figure 5.15: Boxplots of estimates of $TG_{STD,Z}(t)$ at Percentiles of OS ($\tau = 0.8$): TTP, Clayton (top row) and Gumbel (bottom row) Data Generation, Total Gain Application

censoring, with censoring also leading to slightly higher estimates in the later tail of the survival distribution, where few patients remain alive. Between the Clayton (top row) and Gumbel (bottom row) generated datasets the pattern over time is broadly consistent, with the exception that the values based on Gumbel generated data appear to be slightly higher at the earlier end of the survival distribution, likely due to the stronger association between early event times assumed by this model. This means that the increase in values

over time is reduced and the overall pattern is relatively stable.

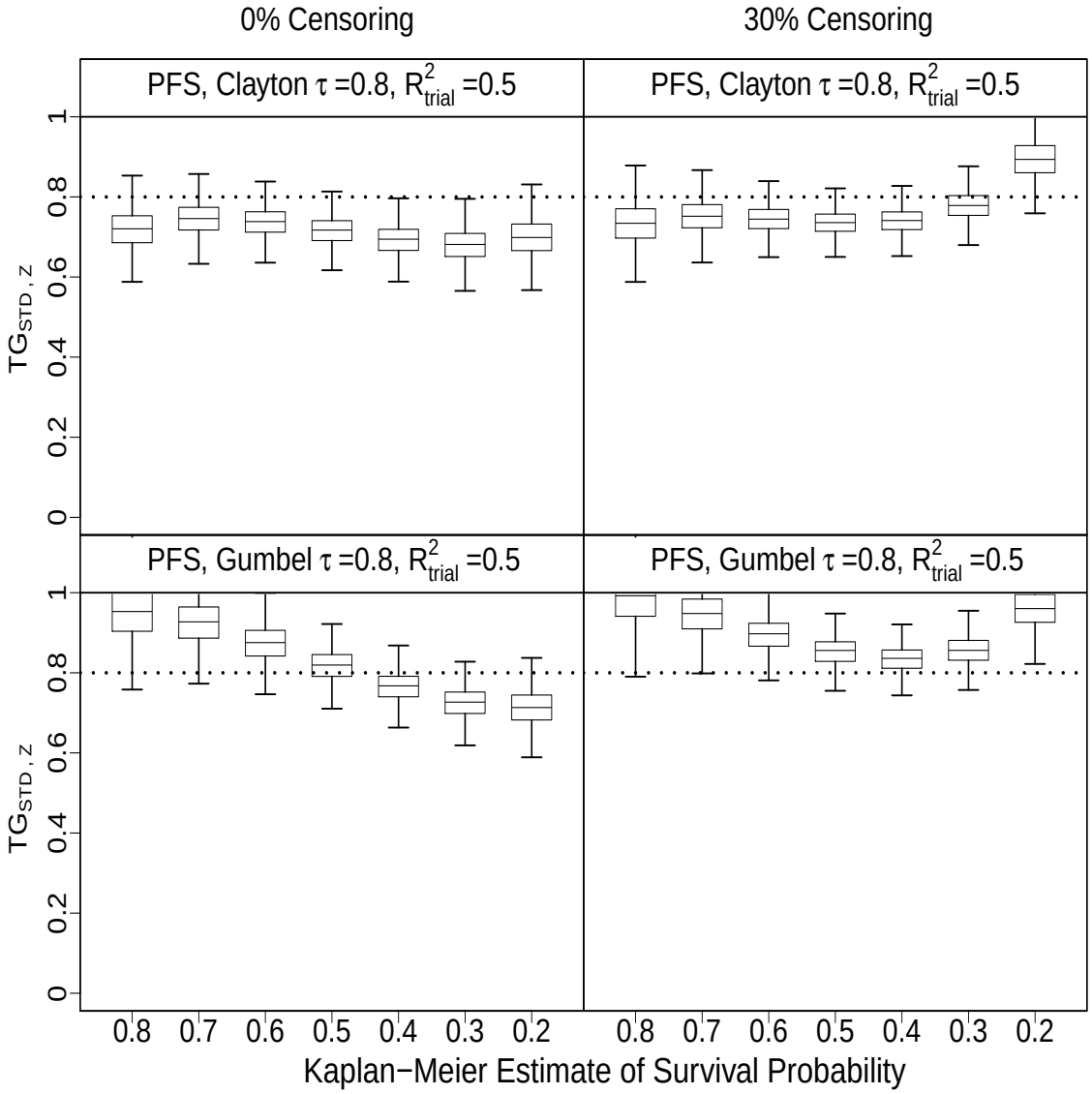

Figure 5.16: Boxplots of estimates of $TG_{STD,Z}(t)$ at Percentiles of OS ($\tau = 0.8$): PFS, Clayton (top row) and Gumbel (bottom row) Data Generation, Total Gain Application

When considering PFS as the potential surrogate endpoint (Figure 5.16), the Clayton generated data (top row) demonstrate a pattern that appears to differ from the setting of TTP. Firstly, the values based on PFS are generally higher and less susceptible to the increase over time, with estimates of $TG_{STD,Z}(t)$ being close to the intended strengths of surrogacy (see Appendix Figure C.3 for lower levels of τ). Values of $TG_{STD,Z}(t)$ appear

to fluctuate only in the extreme tails of the Kaplan-Meier distribution, with estimates across the middle range of the Kaplan-Meier distribution being reasonably constant. As for previous scenarios, the variability in estimates appears to increase slightly in the tails of the Kaplan-Meier distribution, as could be expected, and the presence of censoring leads to slightly higher values in the later survival proportion.

Further investigation of Gumbel copula generated PFS data shows that the change in dependence structure has a similar effect on PFS as was observed for TTP when the true association is low to medium (Appendix Figure C.4). For the highest level of association, increases in the estimates of $TG_{STD,Z}(t)$ in the early tail of the Kaplan-Meier distribution lead to values that are slightly over-estimating the τ value, whereas values in the later tail remain similar to the Clayton generated data. The impact of this change in only the lower portion of the survival distribution leads to patterns over time that look quite different to the other settings. However, in the middle of the survival distribution, the estimates are close to the reference value of τ. In the Gumbel generated data, there are also a number of settings where estimates of $TG_{STD,Z}(t)$ exceed a value of one, and this will be further discussed later in this section.

Overall, the values of $TG_{STD,Z}(t)$ over time lead to three conclusions; that predictive ability of PFS is generally higher than for TTP, that Gumbel generated data provides values higher than Clayton generated data, particularly in the earlier tails, and that censoring has minimal impact on the values, leading to small increases only in the upper extreme of the Kaplan-Meier distribution. These three elements will be discussed further next, in the context of examining the patterns of $TG_{STD,Z}(t)$ over time, the increase in variability in the tails, and the estimates that exceed a value of one.

Further Investigation: Changes in $TG_{STD,Z}(t)$ over time

In order to examine more closely the pattern of $TG_{STD,Z}(t)$ over time, the individual components of the measure were considered. Estimated values of $TG_Z(t)$, $TG_Z(t)_{max}$ and

$TG_{STD,Z}(t)$ for the setting of strongest surrogacy ($\tau = 0.8$, without censoring) are shown in Figure 5.17, with TTP Clayton in the top left, TTP Gumbel in the top right, PFS Clayton bottom left and PFS Gumbel bottom right of the figure. The value of $\tau = 0.8$ was selected as this reflects the extremes of the observed patterns in $TG_{STD,Z}(t)$ over time. Each individual boxplot represents the values across the 5,000 repetitions for a given percentile of the Kaplan-Meier distribution, with the decrease along the x-axis from left to right reflecting the progression of time and therefore reduction in the Kaplan-Meier estimate of survival.

The patterns of values of $TG_Z(t)_{max}$ and $TG_Z(t)$, with the decreases towards the tails of the Kaplan-Meier distribution, are explained by the underlying concept of Total Gain. In the early tail, the vast majority of patients remain event-free and so knowledge of the surrogate outcome offers limited improvement in the prediction of survival status. In the upper tail, the vast majority of patients have experienced the event, so predicted survival is close to zero for all remaining patients and is improved only minimally through knowledge of the surrogate outcome. The greatest gain from knowledge of the surrogate then occurs in the middle range of the Kaplan-Meier distribution, reflected in the shapes in Figure 5.17 and in the work of Choodari-Oskooei et al. (2015), who concluded that the difference between the predicted and average survival probabilities was minimal near the time origin and at the latest timepoints of the survival distribution.

These patterns in $TG_Z(t)_{max}$ and $TG_Z(t)$ then provide further information about the observed patterns in $TG_{STD,Z}(t)$, where slight differences in values of $TG_Z(t)$ in the tails are magnified when scaled using the maximum values. This is further illustrated in Table 5.3, which shows the average (mean) values of $TG_Z(t)$ and $TG_{STD,Z}(t)$ across the 5,000 simulation runs for each scenario presented in Figure 5.17.

It can be seen from this table that the values of $TG_Z(t)$ are always higher for Gumbel data as compared to Clayton data, and are always higher for PFS than for TTP. Both of these findings are consistent with the previous results presented in this thesis. Of interest here, however, is how the estimates change over time. Firstly, the difference in $TG_Z(t)$

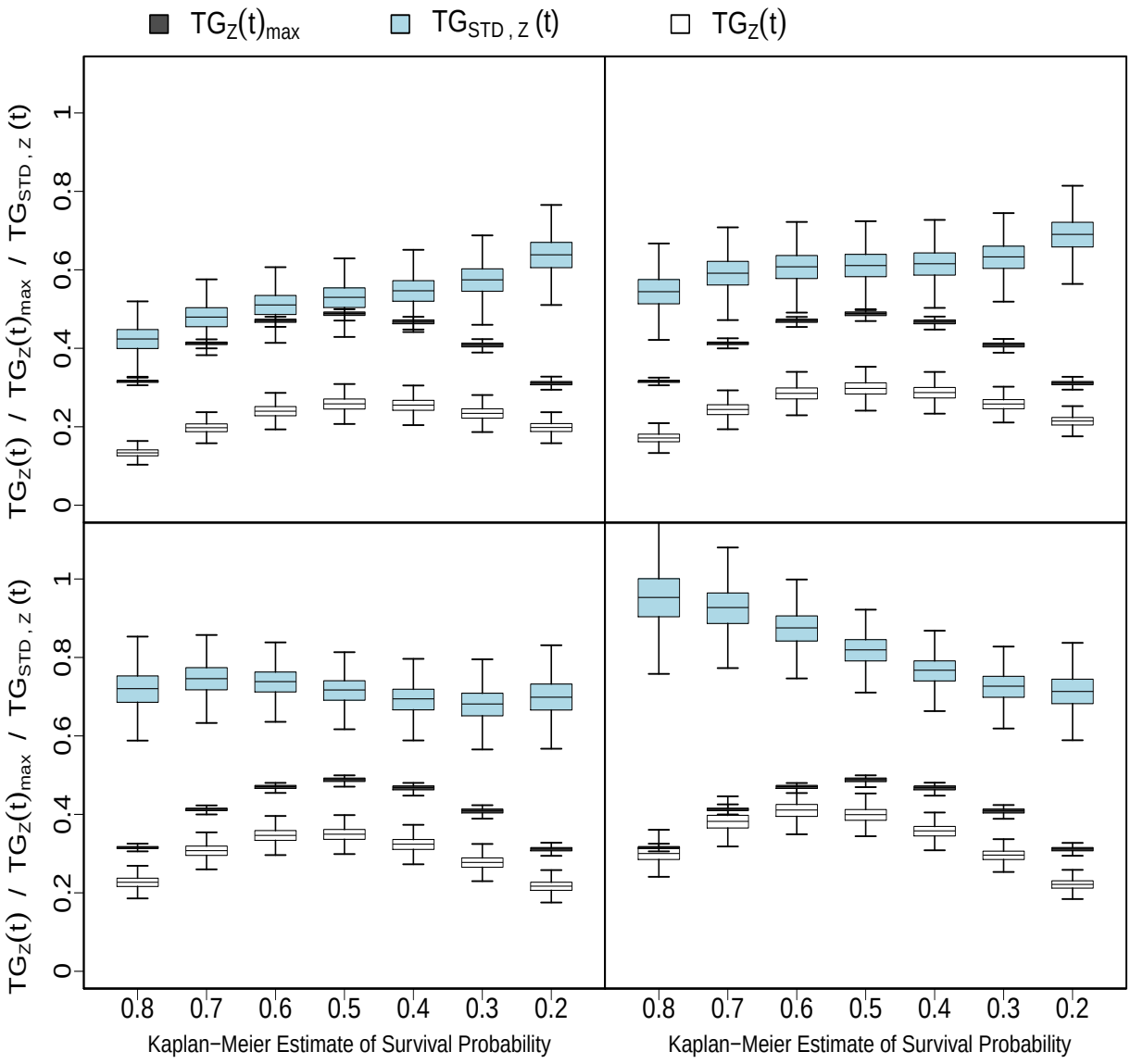

Figure 5.17: Boxplots of estimates of $TG_Z(t)$, $TG_Z(t)_{max}$ and $TG_{STD,Z}(t)$ across the Kaplan-Meier distribution for OS: TTP Clayton (top left), TTP Gumbel (top right), PFS Clayton (bottom left), PFS Gumbel (bottom right) ($\tau = 0.8$, No Censoring)

estimates between Clayton and Gumbel generated data is greater at the earlier Kaplan-Meier percentiles than at the later percentiles, for both TTP and PFS. This reflects the dependence structure of the Gumbel copula, which exhibits stronger dependence between early event times, overall leading to values of $TG_{STD,Z}(t)$ that are higher for the Gumbel data than the Clayton data during the earlier Kaplan-Meier percentiles, as observed in

Figure 5.17 (right hand column compared to left hand column). Secondly, the higher values of $TG_Z(t)$ for PFS as compared to TTP (for both data generation algorithms) leads to values of $TG_{STD,Z}(t)$ that are higher for PFS (bottom row of Figure 5.17) as compared to TTP (top row of Figure 5.17).

Table 5.3: Values of $TG_Z(t)$ and $TG_{STD,Z}(t)$ for $\tau = 0.8$, no censoring

	$TG_Z(t)$ ($TG_{STD,Z}(t)$)			
Kaplan-Meier Percentile	TTP Clayton	TTP Gumbel	PFS Clayton	PFS Gumbel
80%	0.134 (0.424)	0.172 (0.545)	0.227 (0.720)	0.300 (0.952)
70%	0.198 (0.479)	0.244 (0.591)	0.307 (0.745)	0.381 (0.924)
60%	0.240 (0.510)	0.285 (0.606)	0.346 (0.737)	0.410 (0.872)
50%	0.258 (0.529)	0.297 (0.609)	0.348 (0.715)	0.398 (0.816)
40%	0.255 (0.529)	0.286 (0.613)	0.323 (0.692)	0.356 (0.764)
30%	0.233 (0.573)	0.257 (0.631)	0.277 (0.679)	0.295 (0.724)
20%	0.198 (0.637)	0.214 (0.688)	0.217 (0.698)	0.221 (0.712)

These findings are therefore not unexpected, but highlight the need for careful consideration of t, and in particular demonstrate that t must be selected not too close to the tails of the Kaplan-Meier distribution, where estimates of parameters can be based on very few events or on very few patients remaining at risk, leading to instability in modelling. Further, the pattern of dependence of the underlying data structure must be considered, since estimation of $TG_{STD,Z}(t)$ may be biased if based on timepoints that reflect only part of the association structure.

It should be noted that in the extreme tails of the Kaplan-Meier distribution, the data are also unlikely to be reliable. The 20th percentile of the Kaplan-Meier distribution is likely to occur very early in a clinical trial, when the data are not yet considered mature. Based on the small sample sizes investigated here, there will be very few events on which the parameters of interest are calculated, which may make these parameters unstable, and in turn affect estimation of $TG_{STD,Z}(t)$. Towards the end of the Kaplan-Meier distribution,

there are few patients remaining at risk, as well as many other potential confounding factors affecting the analysis. It is therefore advisable to focus estimation on the middle range of the Kaplan-Meier distribution, for example between the $40-60\%$ range, to ensure that modelling parameters can be reliably estimated and are stable. For the majority of scenarios investigated here, estimation of surrogacy within this interval leads to estimates of $TG_{STD,Z}(t)$ that are generally in line with the intended strength of surrogate endpoint, with reasonably small ranges of estimates given the low sample sizes.

Variability in $TG_{STD,Z}(t)$ in the extremes of the Kaplan-Meier distribution

The aforementioned pattern in $TG_Z(t)$ is also considered to help explain the increases in variability that are sometimes observed in the extremes of the Kaplan-Meier distribution. The estimation of model parameters in these extremes is often based on small subgroups of patients, either due to low numbers of events or due to low numbers of patients remaining at risk of an event. With the variability inherent in each of the simulated clinical trial datasets, this can lead to slightly increased variability in estimates of $TG_Z(t)$ as well as $p_1(t)$ and $p_2(t)$, which in turn leads to an increase in the ranges of estimates of $TG_{STD,Z}(t)$. In the majority of scenarios examined, the increase in variability in the tails is mild, and is certainly less of a concern than the increases in estimates discussed above. Importantly, in the mid-range of the Kaplan-Meier distribution, where the estimates are more stable, the change in variability is negligible and does not impact interpretation of results.

Values of $TG_{STD,Z}(t)$ extending above one

The predictive ability of a set of covariates can theoretically only range between a value of zero and one, where zero indicates that the covariates have no value in predicting the outcome of interest and one indicates perfect prediction. The maximum value of $TG_Z(t)$ is intended to reflect this perfect scenario. From Figure 5.17, it can be seen that on occasion the value of $TG_Z(t)$ is overlapping with the maximum value, leading to values of

$TG_{STD,Z}(t)$ that are greater than one when the true underlying $\tau = 0.8$.

This occurs most prominently in the extreme tails of the Kaplan-Meier distribution, where there are fewer than 20% of patients remaining at risk or where only 20% of patients have experienced the event. The over-estimation is considered to be due to the instability of estimates in these extreme tails. When using PFS as the potential surrogate endpoint with Gumbel generated data, the strength of association is at the highest point during the early event times, and so the effect of this covariate overwhelms the treatment effect in the Cox model. This leads to predicted survival probabilities that are close to one for patients without the surrogate outcome and close to zero after the surrogate outcome is observed, regardless of the treatment effect. The effect of this large difference is that in the early tail, the Kaplan-Meier estimate cannot distinguish between these extreme groups, and the area between the two probabilities becomes very large.

Similarly, in the later tail, the estimates are impacted by the level of censoring in the data. When the sample size of 120 patients is further reduced by censoring, there are only a very small number of patients who remain in the risk-set in the later Kaplan-Meier tails. These small numbers, possibly fewer than 10 patients per treatment group, leads to very unstable estimates from both the Cox and Kaplan-Meier models, and the substantial amount of uncertainty in model parameters leads to values of $TG_{STD,Z}(t)$ that are not stable, and should not be considered reliable.

Overall, these additional investigations have demonstrated that the choice of t is critical to ensure that $TG_{STD,Z}(t)$ is stable and reliable enough to make robust conclusions. The sensitivity analyses over the time course of the Kaplan-Meier distribution have demonstrated that while there is variability, and uncertainty when calculating $TG_{STD,Z}(t)$ using the extremes of the available data, the estimates based on the middle of the Kaplan-Meier distribution remain relatively stable.

5.6.3 Larger Treatment Effects

The predictive accuracy of a surrogate endpoint in the newly developed $TG_{STD,Z}(t)$ measure is calculated by determining the difference between the predicted survival probability from a model containing treatment and the surrogate outcome as covariates, and the average survival probability from all patients within each respective treatment group. The magnitude of treatment benefit observed on the surrogate and true endpoints is therefore a key factor, since the intention is to assess the predictive ability of the surrogate after having accounted for the observed treatment effect.

Throughout this chapter, simulation studies have focused on one fixed value of the treatment effect for each endpoint, with hazard ratios of ≈ 0.67 for PFS and ≈ 0.82 for OS. To examine the impact that changes in treatment effects can have on estimation of $TG_{STD,Z}(t)$, additional examination of larger treatment effects have also been considered. The aim of such exploration is to determine whether the change in treatment effect has any impact on the calculation of predicted survival probabilities that would confound the estimate of predictive ability of the proposed surrogate. Additional investigation focuses on treatment effects that demonstrate a greater benefit, with hazard ratios of ≈ 0.50 for PFS and ≈ 0.60 for OS. Results are presented in Figure 5.18 for TTP and Figure 5.19 for PFS, where Clayton copula generated data are displayed in the top row and Gumbel copula generated data in the bottom row.

Given that the data generation procedure defines the strength of surrogacy and strength of treatment effect separately, it is not expected that estimates of $TG_{STD,Z}(t)$ would be sensitive to changes in the treatment effects. Particularly, the newly developed approach accounts for treatment effects on both endpoints when estimating predictive ability. Reassuringly, for all scenarios examined, there appear to be negligible differences in estimates of $TG_{STD,Z}(t)$, across both surrogate endpoints and both data generation methods, as well as in the presence of censoring. Estimates are very close to those based on the original treatment effects, indicating that $TG_{STD,Z}(t)$ is able to appropriately adjust for the ob-

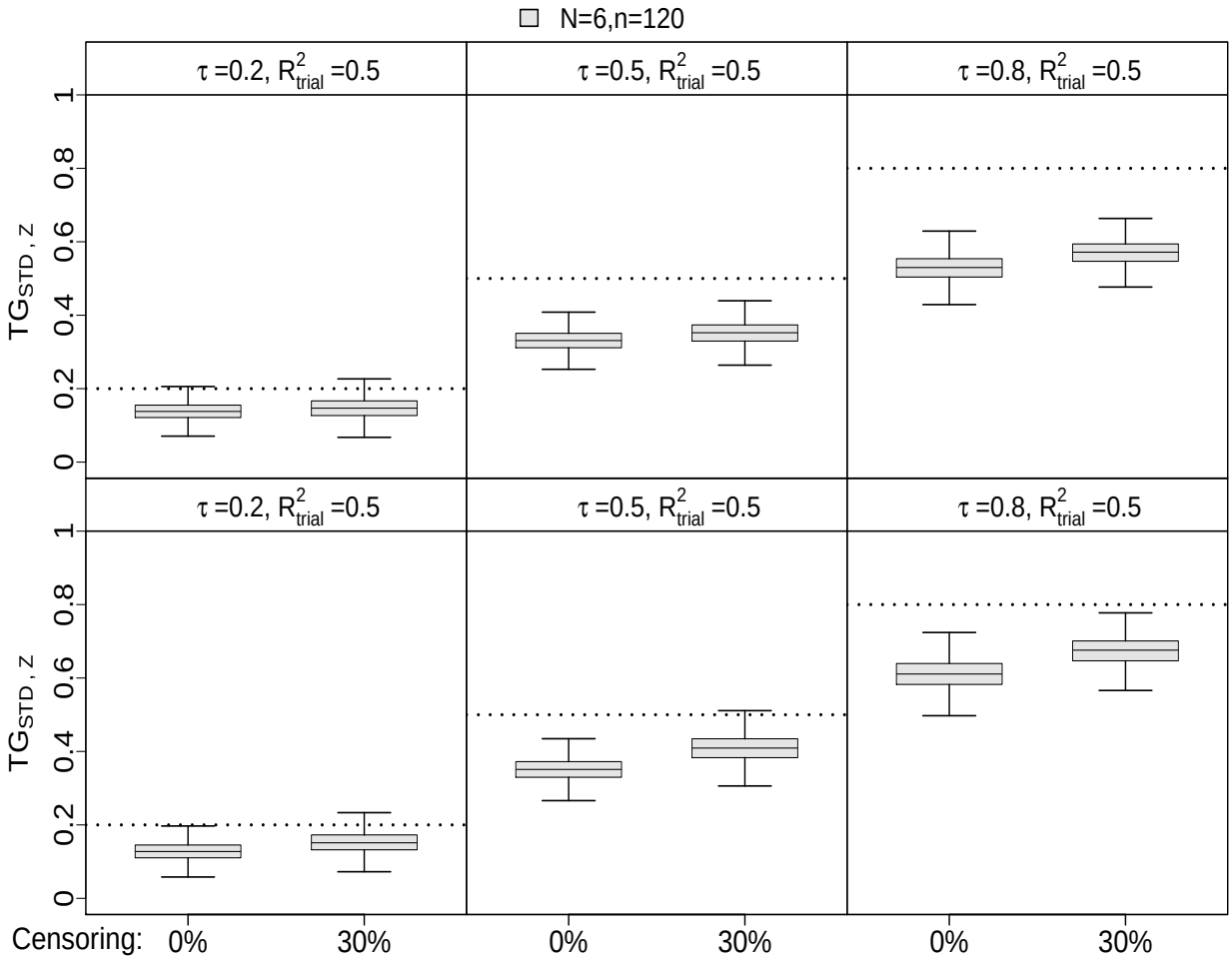

Figure 5.18: Boxplots of estimates of $TG_{STD,Z}(t)$ at Median OS - larger treatment effects: TTP, Clayton (top row) and Gumbel (bottom row) Copula Data Generation, Total Gain Application

served treatment effect and isolate the predictive ability of the surrogate endpoint. This demonstrates that $TG_{STD,Z}(t)$ is not sensitive to external changes in the structure of the datasets that do not affect the underlying strength of surrogacy.

5.6.4 Larger Sample Sizes

Whilst the setting of predominant interest in this thesis is small sample sizes, it is highly relevant to ensure that the newly proposed methodology also performs satisfactorily when

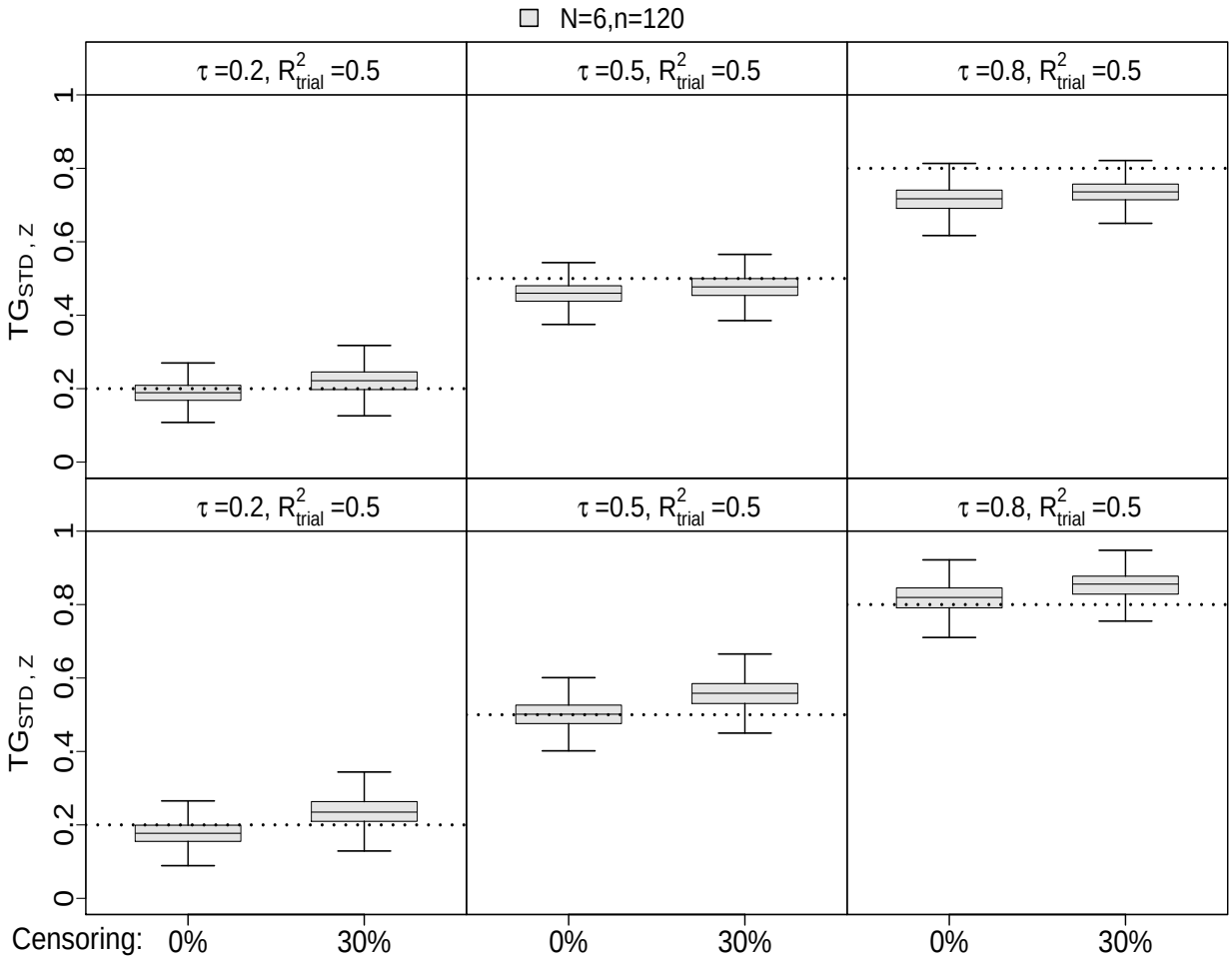

Figure 5.19: Boxplots of estimates of $TG_{STD,Z}(t)$ at Median OS - larger treatment effects: PFS, Clayton (top row) and Gumbel (bottom row) Copula Data Generation, Total Gain Application

more data are available to evaluate potential surrogates, which is the ideal case. Further simulations were therefore conducted to examine how $TG_{STD,Z}(t)$ performs when more and larger clinical studies are available for analysis. In particular, ten clinical trials each containing 500 patients are examined, under no censoring and 30% censoring, for both TTP and PFS. Given the broad similarity of results across previous scenarios, only Clayton generated data are considered. Results are presented in Figure 5.20, with TTP presented in the top row and PFS in the bottom row.

The expectation for these results is that the larger sample sizes will lead to more precise

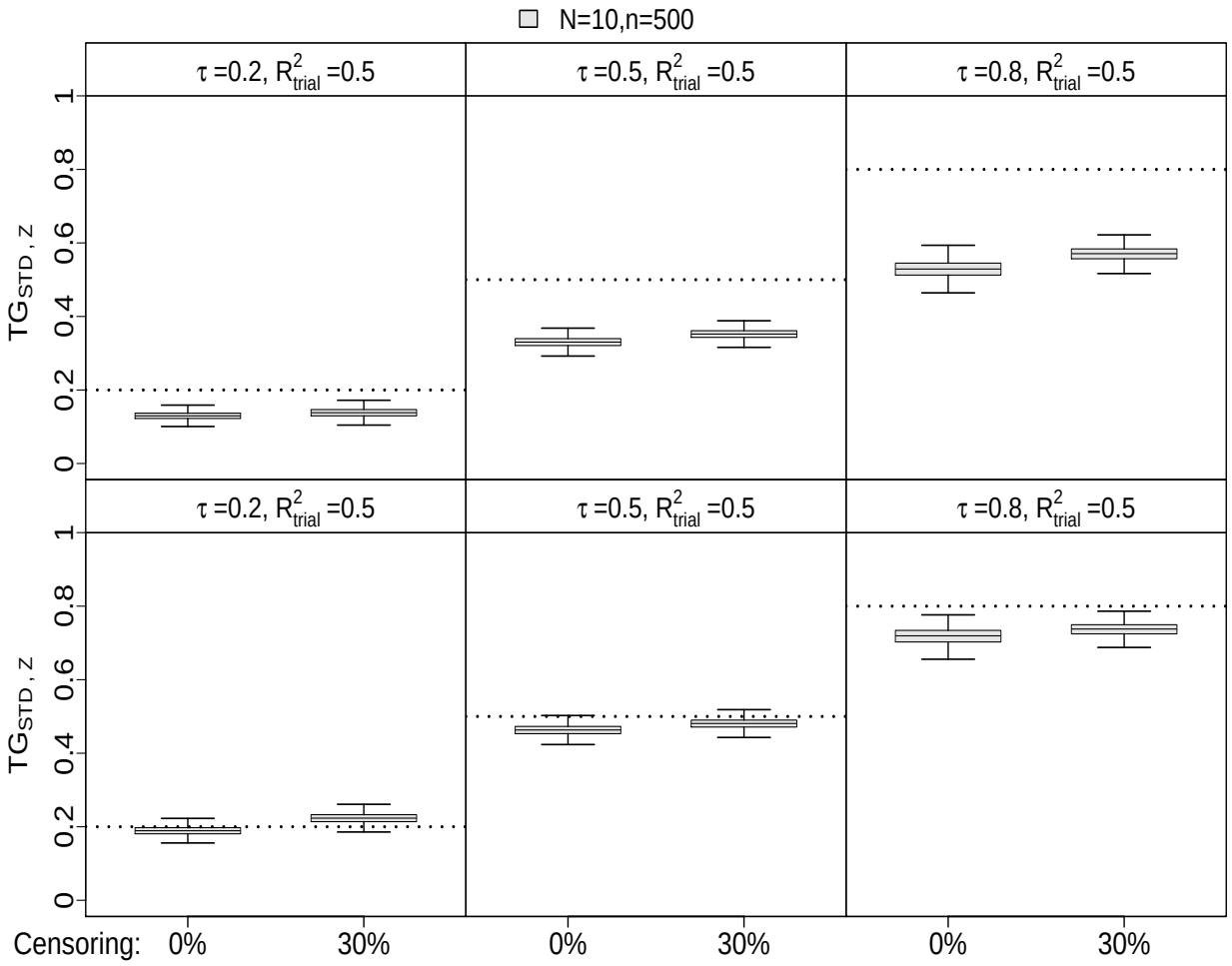

Figure 5.20: Boxplots of estimates of $TG_{STD,Z}(t)$ at Median OS - $N = 10, n = 500$: TTP (top row) and PFS (bottom row), Clayton Copula Data Generation, Total Gain Application

estimation of model parameters, leading to lower variability in values of $TG_{STD,Z}(t)$ across the $5,000$ repetitions as compared to the setting of small sample sizes. This is indeed the case, with smaller ranges of estimates for all of the scenarios investigated and clear separation in ranges across the true underlying strengths of surrogacy. Encouragingly, there is little change in the median estimate for each setting, demonstrating that while inclusion of larger sample sizes has improved the precision, $TG_{STD,Z}(t)$ can also be reliably estimated when based on smaller samples. For the TTP setting, the values of $TG_{STD,Z}(t)$ continue to be estimated lower than the reference value of τ and this will be discussed in

the next section. For PFS, the conclusion is similar; the increase in sample size has led to more precise estimation of $TG_{STD,Z}(t)$, with very narrow ranges of estimates that are clearly separated across values of τ. The under-estimation observed with TTP is not of concern for the setting of PFS, with estimates lying very close to the intended strength of surrogacy. Overall, these additional simulations therefore support the reliability of $TG_{STD,Z}(t)$ as a measure of individual-level surrogacy.

5.7 Understanding the Results

Despite the strong performance of $TG_{STD,Z}(t)$ in evaluating individual-level surrogacy for the scenarios examined, particularly for PFS, there are a number of observations that must be addressed, including the apparent under-estimation of the assumed underlying strength of association between TTP and OS, and the observed variability in estimates, and these are discussed further in this section.

5.7.1 Comparing TTP and PFS

In scenarios based on TTP as the surrogate endpoint, estimates of $TG_{STD,Z}(t)$ were lower than the input value of τ for both Clayton and Gumbel data generation, with the exception of the lowest strength of association. Values based on Gumbel generated datasets were slightly higher than those for the Clayton generated data, however the estimates of $TG_{STD,Z}(t)$ remain under-estimated, as compared to τ, for both data generation approaches. This was not the case for PFS, where estimates were close to the input value of individual-level surrogacy across the majority of scenarios.

In order to understand the differences between results for TTP and PFS, it is necessary to examine the individual components of the calculation of $TG_{STD,Z}(t)$. First, the maximum value of $TG_Z(t)$ is constant regardless of the choice of surrogate endpoint, since this is calculated independent of all covariates. Similarly, the Kaplan-Meier estimates at the time of median OS, $p_1(t)$ and $p_2(t)$ are based only on the covariate of treatment and

are therefore identical in both settings. The only parameter that can introduce differences in the value of $TG_{STD,Z}(t)$ between TTP and PFS is therefore the predicted survival probability from the Cox proportional hazards model. For values of $TG_{STD,Z}(t)$ to be lower for TTP, the predicted survival probability would need to be closer to the reference value coming from the Kaplan-Meier function, such that the predictive ability of TTP is lower than that of PFS. As noted above, this would be expected given that PFS includes the true endpoint of interest.

The values of $TG_{STD,Z}(t)$ that are below the reference value, τ, are therefore considered to be due to the use of the Cox proportional hazards model. Since this model does not estimate τ, but rather estimates a covariate coefficient that does not have a fixed range, it is not expected that values of the two would match exactly. Whilst the underestimation is present, the results have demonstrated that $TG_{STD,Z}(t)$ provides estimates of individual-level surrogacy that can adequately differentiate poor, mediocre and strong surrogates, even if the exact values of the measure do not match a theoretical threshold. The underestimation is therefore considered a slight limitation of the approach when based on TTP, however the simplicity of computation and interpretation provide good rationale for use in practice.

5.7.2 Variability

Results of $TG_{STD,Z}(t)$ presented in Section 5.5.2 demonstrate that while the ranges of estimates have only minimal overlap between true levels of association, the variability increases slightly with increased individual-level association, particularly for TTP. This is consistent with the information theory approach, although to a much lesser extent. Since both the information theory approach and $TG_{STD,Z}(t)$ use the same underlying models, it is expected that there is a consistent reason for the increased variability. In Section 4.4.3, it is noted that as the underlying strength of surrogacy increases, the variability in the estimated coefficient of the surrogate time-dependent covariate in the Cox proportional

hazards model also increases (see also the left hand side of Figure 4.11). The predicted survival probability from this model that is used in estimation of $TG_{STD,Z}(t)$ is therefore also subject to this increase in variability, leading to wider ranges of values when based on increased values of τ. This is further supported through the results based on increased sample size, where the greater data availability leads to more precise estimates of Cox model parameters, and subsequently more consistent estimates of $TG_{STD,Z}(t)$. Importantly, the variability observed in $TG_{STD,Z}(t)$ does not hamper conclusions, and it is possible to reliably identify truly strong surrogate endpoints.

5.8 Implications of Results

Previous sections of this chapter have presented a wide range of simulation results, including multiple sensitivity analyses, to ensure that the proposed methodology is robust and reliable. In order to put these findings into further context, additional discussion is provided herein.

5.8.1 Practical Implications

The novel measure presented in this chapter offers many benefits for the evaluation of individual-level surrogacy. Firstly, it can be calculated very simply using standard statistical software that is employed in the analysis of clinical trials. For time-to-event data, the use of Cox proportional hazards and Kaplan-Meier models is commonplace, and model parameters required for estimation of $TG_{STD,Z}(t)$ are those that are already used commonly to make inferences about survival data. This concept is appealing to statisticians, but also when discussing results with clinicians, who are often familiar with the terminology and concepts of these standard approaches. Furthermore, the lack of distributional assumptions of these models (beyond that of proportional hazards) make them applicable to a wide range of clinical trial databases, widening the scope of the new approach without requiring modification to the estimation process. Indeed, as mentioned previously,

the method is also applicable to a wide range of alternative models that can be used to estimate the predicted survival probability after accounting for covariates, including parametric models. This ease of calculation is further reflected in the ease of interpretation of the approach, as the ability of the surrogate endpoint to predict the true endpoint after accounting for treatment (and other covariates). This concept is highly relevant, and easy to explain to non-statisticians, particularly with the use of the graphical representation. As compared to the two-stage meta-analytic copula approach, the lack of joint modelling of the surrogate and true endpoints also improves the applicability in practice, as there is no need for such complex modelling with associated challenges.

More specifically, related to the results of simulation studies, the $TG_{STD,Z}(t)$ approach has been shown to be highly effective in reflecting varied strengths of surrogacy in a setting where PFS is considered as the potential surrogate endpoint, a setting where both of the previously examined methods in this thesis have demonstrated poor performance. Whilst performance in the setting of TTP demonstrated under-estimation, PFS is arguably the more relevant endpoint. By the very nature of surrogacy, the aim is to identify endpoints that can mature sooner, such that long clinical trials are not required to establish the clinical benefit of a new therapeutic agent. PFS includes both disease progression and death as events of interest, and so these events accumulate faster than for TTP, which considers only disease progression of interest. For diseases where patients may die before they experience diagnosed disease progression, this can make a notable difference to the requirements for follow-up and sample sizes within clinical trials. More importantly, TTP cannot capture the impact of treatment on the event of death, which is a highly relevant outcome given that it is considered the true endpoint of interest. In addition, PFS is a well understood and recognised endpoint by statisticians, clinicians and health authorities, and has been used as the basis for a number of regulatory approvals for new treatments. Interestingly, a proposal for surrogacy evaluation based on separation of PFS into the individual events of disease progression and death attracted a lengthy discussion around the strong relevance of PFS as a composite endpoint, and the need to keep the two events

of interest in this composite form (Ghosh et al., 2012).

For diseases where TTP is considered the more relevant endpoint, use of both the two-stage meta-analytic copula and $TG_{STD,Z}(t)$ approaches is recommended, with the two-stage meta-analytic copula method having strongest performance under correctly specified models. Overall, the strong performance of $TG_{STD,Z}(t)$ in estimating individual-level surrogacy of PFS (for OS), alongside the aforementioned benefits of the approach, support that the new methodology is reliable and can be recommended for this use in practice.

5.8.2 Limitations of the Simulation Study

The newly proposed method, $TG_{STD,Z}(t)$ has been studied using simulated clinical trial datasets for a wide range of scenarios, including contrasting underlying data structures, varied strength of association between surrogate and true endpoints, different surrogate endpoints and under censoring. Additional sensitivity analyses examined an alternative data generation algorithm, changes in treatment effects observed on both endpoints, variation in the timing of the evaluation of $TG_{STD,Z}(t)$ and larger sample sizes. Despite this thorough investigation, the simulation studies are subject to limitations.

As was the case for the information theory method, the primary limitation is that in the simulation study the data were not generated to a specific strength of individual-level surrogacy as defined by the method being investigated. Since $TG_{STD,Z}(t)$ is based on estimation of parameters from multiple different models, all of which adjust for covariates, it is difficult to accurately control the level of association between surrogate and true endpoints. To ensure that each sample is being compared to a constant reference value, and to ensure informal comparability between surrogacy evaluation methods is possible, a copula model was used to generate simulated datasets. The impact of this is that the true underlying value of the (transformed) copula parameter τ may not accurately reflect the true value that is being estimated by $TG_{STD,Z}(t)$. However, consideration of two different copula models, as well as sensitivity analysis not based on copula modelling, demonstrated

consistent results and show that $TG_{STD,Z}(t)$ is able to reliably detect low, medium and high strengths of surrogacy. Whilst these strengths may not be exactly equal to values of 0.2, 0.5 or 0.8 respectively, the results are within a reasonable region of these input values, and conclusions drawn from estimated values of $TG_{STD,Z}(t)$ are not considered to be hampered. Further investigation of alternative data generation procedures is a topic for future research.

A further limitation is that the $TG_{STD,Z}(t)$ measure is based on a model that makes an assumption of proportional hazards, such that the covariates have a multiplicative effect on the hazard function. No consideration was taken for the impact on estimation when this assumption is violated. However, examination of the original measure $TG_{STD}(t)$ under non-proportional hazards for various values of t was conducted by Choodari-Oskooei et al. (2015). As an example, Choodari-Oskooei et al. (2015) consider a covariate of treatment only, where survival curves cross over partway through the survival distribution such that the control arm has superior survival for the first portion of the study and the experimental arm has superior survival after the time of the curves crossing. Results of $TG_{STD}(t)$ over time demonstrated that the method reflects this by tending towards zero when approaching the time at which the survival curves cross, and increasing thereafter. Therefore, as previously discussed, the time at which $TG_{STD,Z}(t)$ is calculated is of critical importance and needs to be carefully selected.

5.9 Further Work

The concept of Total Gain in the assessment of baseline covariates was introduced by Bura and Gastwirth (2001) in the setting of binary outcomes, and extended to build prognostic models in a survival setting by Choodari-Oskooei et al. (2015). The development of $TG_{STD,Z}(t)$ as a measure of surrogacy within this thesis has further extended the method within the survival setting, however the underlying methodology is applicable to many different endpoint types. The fundamental concept of Total Gain is based on estimation

of two predicted probabilities, one based on an average for the sample under study, and one based on the individual combination of covariate values and respective coefficients from a selected model. Therefore, any setting that allows such models to be built will provide the quantities needed to calculate the measure. For extension to other settings, the derivation of the maximum value is needed, however the estimation of predictive ability of a set of one or more covariates remains unchanged. Therefore, further extension to continuous, categorical or longitudinal endpoint types is worthy of further examination.

With regards to the survival setting, the current investigation has been conducted using Cox proportional hazards models and Kaplan-Meier survival estimates in the calculation of Total Gain, as was that of Choodari-Oskooei et al. (2015). Further work could therefore be considered to understand what level of proportional hazards violation would lead to a deterioration in performance of $TG_{STD,Z}(t)$. Additionally, alternatives to the Cox proportional hazards model could be considered in estimation of the predicted survival probability.

Specific to the context of surrogate endpoint evaluation, a natural extension would be to consider the Total Gain concept in the assessment of trial-level surrogacy. Rather than quantifying the ability of the surrogate endpoint to predict the outcome of the true endpoint, it would be of interest to quantify the ability of treatment effect on the surrogate endpoint to predict treatment effect on the true endpoint. In such a setting, the treatment effect on the true endpoint would be modelled with consideration of the treatment effect on the surrogate endpoint as a covariate, as compared to the average treatment effect on the true endpoint across all studies. In previous chapters, a linear relationship between treatment effects on the two endpoints is assumed, and so this could be further considered.

Finally, the use of Total Gain in the evaluation of surrogate endpoints has previously been considered by Huang and Gilbert (2011), however this was in the framework of principal surrogacy, as introduced in Section 2.6.3. Given that such methodology is considered an emerging area of research (Ensor et al., 2016), further work could be conducted to investigate the Total Gain concept within this alternative framework of surrogacy methodology.

Chapter 6

Illustrative Example: A Phase III Clinical Trial in Gastric Cancer

6.1 Introduction

In order to illustrate the use of all three of the previously described surrogacy methods in practice, and see how they compare when applied to the same real-life dataset, a case study is provided herein. This example was selected as being a close match to the assumptions of the simulated datasets of previous sections, with median PFS of approximately six months and median OS of approximately 12 months.

The dataset originates from the ToGA (Trastuzumab for Gastric Cancer) clinical trial (Bang et al., 2010), which was an international phase III randomised controlled trial undertaken in 122 centres in 24 countries. Patients with gastric or gastro-oesophageal junction cancer were randomised to receive chemotherapy only or chemotherapy in combination with the HER2 targeted therapy, trastuzumab, and the trial had a primary endpoint of overall survival, with PFS measured as a secondary endpoint.

A total of 594 patients were randomly assigned to one of the two study treatments, of whom 584 were included in the primary analysis. An interim analysis of OS was performed after 75% of the required survival events had been observed, and at this time the median

OS was 13.8 versus 11.1 months in the experimental (trastuzumab-containing) and control (no trastuzumab) treatment arms respectively. The hazard ratio of 0.74 (95% confidence interval [0.60 0.91]) was sufficient to cross the pre-specified interim stopping boundary, triggering full reporting of the trial. The PFS result was consistent with OS, demonstrating evidence of a statistically significant benefit from treatment with experimental therapy (median PFS 5.5 months versus 6.7 months, hazard ratio 0.71 [95% confidence interval 0.59 0.85]). Censoring proportions for PFS and OS are 21% and 41%, respectively.

6.2 Modelling Assumptions

In order to assess the performance of PFS as a surrogate endpoint for OS based on the ToGA database, and in particular to align with the previous considerations of this thesis, this large phase III study was split into subgroups, based on the geographical location of the patients. Whilst it has been noted earlier in the thesis that such an approach may not always be the most appropriate way to conduct surrogacy analysis (Renfro et al., 2014), the large sample size from this trial allows for separation into groups of a similar size to those used in the simulation studies of previous chapters, thereby reflecting the setting of interest for this research.

The country in which a patient was treated was selected as the grouping factor, and where individual countries enrolled only a small number of patients, they were combined with other countries from a similar geographical region of the world wherever possible, to ensure that each individual group had at least one patient and one event per treatment arm. Two countries were removed from the analysis due to small numbers and the absence of a geographically similar country to combine with (South Africa with n=4 and Turkey with n=6 patients), leaving 574 patients available for analysis across eight subgroups (compared to ranges of 320-720 patients across 4-6 trials used in the simulation studies). The final groups are summarised in Table 6.1 below.

Table 6.1: Subgroups Used in the Analysis

Group (Countries)	Number of patients
Portugal, UK, Italy, France, Belgium, Spain, Germany, Denmark, Finland	125
Brazil, Peru, Panama, Mexico, Costa Rica, Guatemala	52
China and Taiwan	86
Japan	101
Korea	122
India	10
Australia	13
Russia	65

The endpoints of PFS and OS within this study are defined in the same way as for the simulation studies presented in earlier chapters of this thesis, and are unchanged from previous reporting of this study (Bang et al., 2010):

- **Progression-Free Survival (PFS)**, defined as the time from randomisation until the patient experiences disease progression or death, whichever occurs first. Patients who do not experience disease progression or death during the period of observation are censored at the time that their disease was last assessed by the treating physician.
- **Overall Survival (OS)**, defined as the time from randomisation until death. Patients who remain alive at the end of follow-up are censored at the time they were last known to be alive.

Since the two-stage meta-analytic copula and information theory approaches performed poorly in estimation of trial-level surrogacy when explored via simulation, and considering that the Total Gain measure is not currently developed to assess trial-level surrogacy, the illustrative example is limited to individual-level surrogacy only.

6.3 Results

Estimates of individual-level surrogacy for each of the two-stage meta-analytic copula, information theory and Total Gain methods are provided in Table 6.2. For the Total Gain measure, three timepoints have been selected at 10, 12.5 and 15.4 months, corresponding to the times at which 60%, 50% and 40% of patients remained alive, respectively, according to the Kaplan Meier estimates for OS.

Table 6.2: Individual-Level Surrogacy Estimates for ToGA

Method	Individual-Level Surrogacy (95% Confidence Interval)
Copula	$0.66\ (0.62 - 0.70)$
Information Theory	$0.56\ (0.38 - 0.68)$
Total Gain ($TG_{STD,Z}(10)$)	$0.78\ (0.76 - 0.94)$
Total Gain ($TG_{STD,Z}(12.5)$)	$0.81\ (0.75 - 0.89)$
Total Gain ($TG_{STD,Z}(15.4)$)	$0.94\ (0.78 - 0.99)$

As can be seen from the results, individual-level surrogacy estimates can vary quite widely, with values as low as 0.56 based on the information theory approach, and as high as 0.94 from the largest of the Total Gain estimates. The estimate based on the two-stage meta-analytic copula method lies between these estimates.

It is important to interpret these results in the context of the simulation studies presented in earlier chapters of this thesis. First, it has been highlighted that the dependence strucuture of the observed data can influence the two-stage meta-analytic copula method, and so a scatterplot of the observed values of PFS (S) and OS (T) is provided in Figure 6.1. This illustrates that the underlying data exhibit stronger dependence between early event times rather than later event times, supporting that the data are more likely to follow a Gumbel copula model than a Clayton copula model, as discussed in Sections 3.2.4 and 3.2.5. Therefore, results of the surrogacy analysis should be compared to data gener-

ated according to the Gumbel copula or lognormal algorithms, rather than data generated according to the Clayton copula.

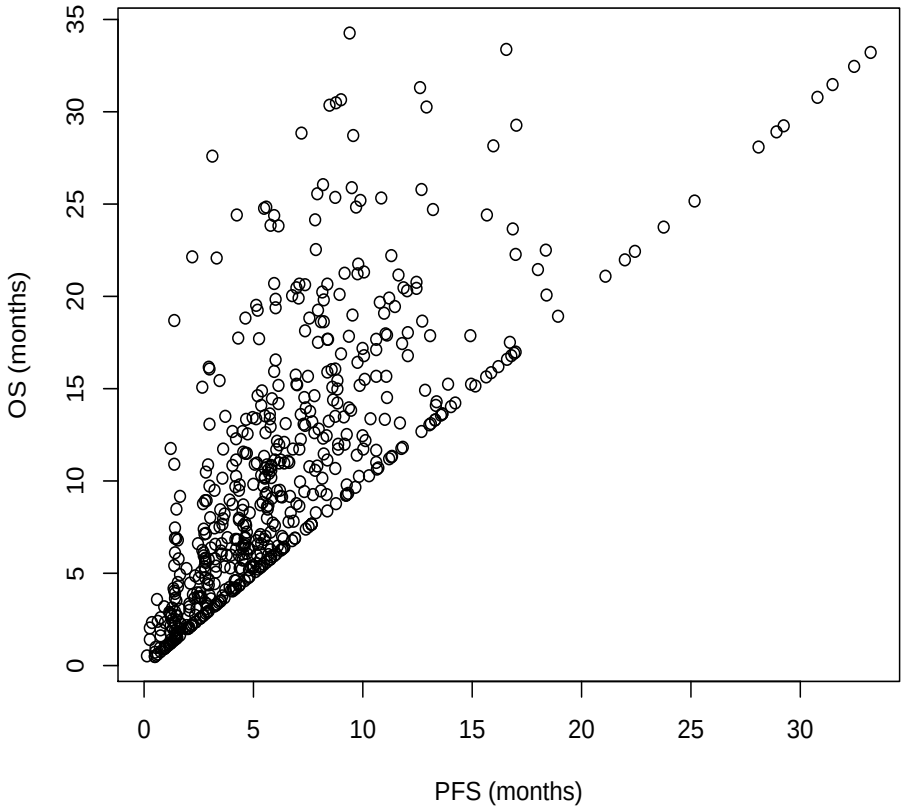

Figure 6.1: Scatterplot of PFS (S) and OS (T) from ToGA

For ease of comparison, the results of the simulation studies for the settings closest to this illustrative example are provided in Figure 6.2 for data generated using the Gumbel copula and Figure 6.3 for data generated using the lognormal algorithm. The closest setting to the real-life example is considered to be the setting of PFS, with 6 trials each containing 120 patients, under 30% censoring. It should be noted that these results are restricted to a true trial-level surrogacy value of 0.5, since results did not appear different across different underlying R^2_{trial} values within the assessments of Chapters 3 and 4.

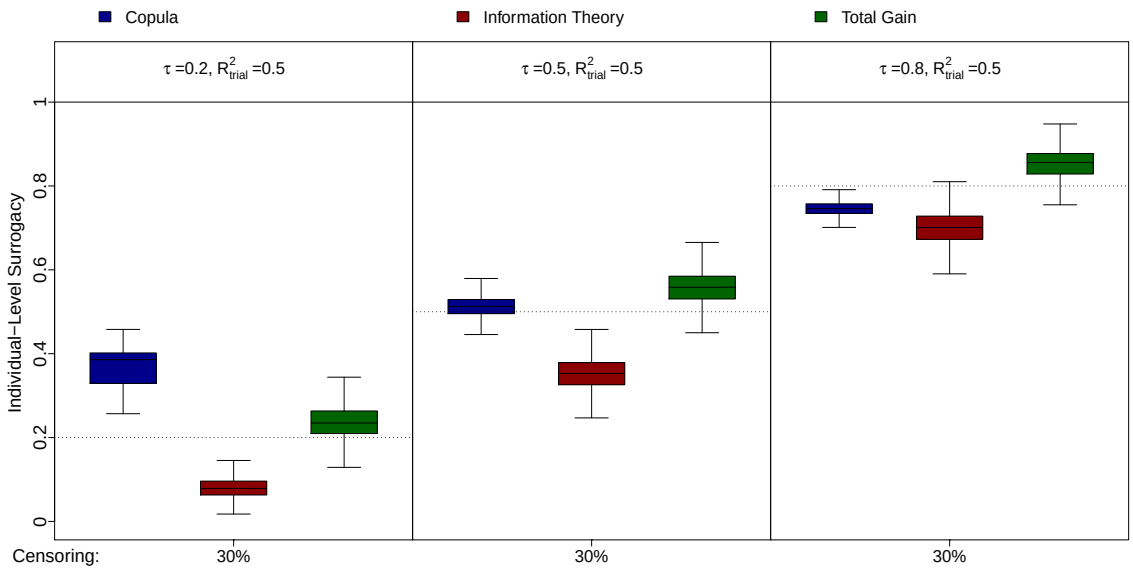

Figure 6.2: Boxplots of all surrogacy methods based on Gumbel copula-generated data ($N = 6$, $n = 120$, 30% censoring)

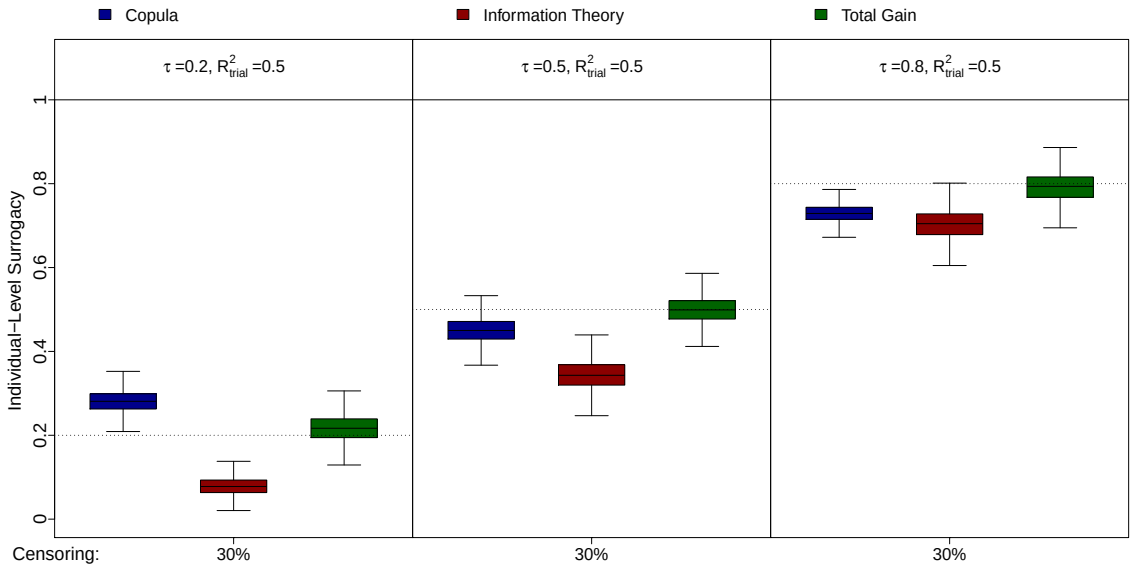

Figure 6.3: Boxplots of all surrogacy methods based on lognormal-generated data ($N = 6$, $n = 120$, 30% censoring)

Based on these previous simulation results, it could be expected that if the true underlying individual-level surrogacy was low (≈ 0.2), then a surrogacy assessment would provide the highest estimate from the two-stage meta-analytic copula, followed by the

Total Gain approach, and with the lowest estimate coming from the information theory measure. When the true underlying surrogacy is medium (≈ 0.5) or high (≈ 0.8), the highest estimate would be expected from the Total Gain method, followed by the copula model, and again with the information theory approach providing the lowest estimates.

Comparing the results from application of the three methodologies to the illustrative example with those presented in Figures 6.2 and 6.3, it is expected that the true underlying individual-level surrogacy of this clinical trial is quite high. The results show the highest estimate for the Total Gain approach ($0.78 - 0.94$), followed by the copula model (0.66) and finally the information theory approach (0.56), and this pattern is broadly in agreement with those observed in the simulations based on the highest level of τ. There is a slightly higher than expected difference between the lowest and highest estimates, but this does not differ substantially from the simulation results, and does not impact the overall interpretation of the result. Use of the two-stage meta-analytic copula approach and the Total Gain approach provide a similar interpretation of strong association between S and T, and support further consideration of PFS as a surrogate for OS in this setting.

These findings further support the research of this thesis by demonstrating that the pattern in observed results between the three surrogate endpoint methodologies is broadly in line with those based on the simulations, and further support that use of the copula models for data generation has not biased the conclusions from those simulations.

Chapter 7

Conclusions and Discussion

7.1 Summary of Research Findings

7.1.1 Review of Original Aims

The use of surrogate endpoints in clinical trials involves a complex decision making process, of which statistical evaluation is a critical component. Prior to being considered appropriate for use by clinical researchers and regulatory authorities, any potential surrogate endpoint must undergo a thorough assessment of the biological plausibility, and the ability to reliably capture information about the true endpoint. A number of statistical methods have been proposed for this purpose, however many questions remain unanswered.

The primary aims of this research were to address these unanswered questions for settings that are of most interest for individual pharmaceutical companies conducting oncology clinical development. In this setting, data may be available from only a small number of previously completed Phase I or Phase II studies within the same, or a similar, clinical development program. These studies may differ in key characteristics that make them unsuitable for meta-analysis, such as differences in the mechanism of action of the control or experimental treatments, a lack of control arm, different patient populations, or varied lengths of treatment or follow-up. Many of the methods proposed for surrogacy

evaluation have been examined under the assumption that multiple, large, clinical trial databases are available. While this is the ideal scenario, it is not considered easy to achieve in practice.

A further aim of this research was to consider surrogate endpoints that are of a composite nature, such that the definition of the surrogate endpoint also includes information directly relating to the true endpoint. In particular, focus has been on oncology settings, where the endpoint of progression-free survival is a well-recognised and understood measure of clinical benefit that has also been used for regulatory approval of new drugs. Whilst this endpoint is mostly oncology specific, the concept of composite endpoints translates to other disease areas, such as cardiovascular trials that aim to assess the treatment effect on composite measures of heart failure and death. The focus herein has therefore been on the evaluation of time-to-event surrogate endpoints for time-to-event true endpoints, where overall survival has been considered the gold-standard true endpoint.

Finally, there is currently no consensus as to which of the methodologies are considered most appropriate for any given setting. Whilst a range of measures have been proposed, many of these are tailored to specific endpoint types, are based on different statistical frameworks, or are measured on different scales. This makes it difficult to determine which methodology may be most appropriate, and whether different methodologies may lead to different conclusions. Examination of the available methodology for the specific setting of interest in this research was therefore necessary to ensure that the most reliable conclusions could be drawn, and to make clear recommendations for future use.

This research addresses these aims by identifying and examining two statistical approaches for the evaluation of surrogate endpoints, to determine their performance in the setting of small sample sizes, to assess the impact of non-symmetric endpoints such as progression-free survival as a surrogate for overall survival, and to identify which of the two may be considered more appropriate for use. Further, a new method for evaluating individual-level surrogacy has been developed and is proposed for future use with time-to-event surrogate and true endpoints.

7.1.2 Trial-level association

Results of the simulation studies conducted in this research have revealed that, as would be expected, large sample sizes are needed to enable reliable estimation of trial-level association. Use of only a small number of Phase I or Phase II trials has been shown to be insufficient for reliable estimates of surrogacy, preventing use of the methods in this setting. Some applications of the methodology have avoided this low number of trials through the use of subsets of trials as the unit for analysis, as discussed by Renfro et al. (2014). However, while this increases the number of data points, it subsequently reduces the number of patients per data point, and so uncertainty in other model parameters is increased. It is recommended that efforts are made to improve the collaboration between multiple clinical research companies, to ensure that sufficient data can be made available to thoroughly evaluate trial-level surrogacy.

7.1.3 Individual-level association

Turning to individual-level surrogacy, the results of simulation studies of the two investigated surrogacy approaches appeared to be more encouraging. In particular, strong performance of the two-stage meta-analytic copula method was observed when time-to-progression was considered the surrogate endpoint and the selected copula was correctly specified. However, when considering the more commonly used endpoint of progression-free survival, performance of the two-stage meta-analytic copula method was poor. The implications of this finding are critical to the evaluation of surrogate endpoints, since use of the two-stage meta-analytic copula method in assessing the relationship between non-symmetric endpoints can lead to erroneous conclusions. In practice, this means that there is a risk for poor surrogates to be confirmed suitable for use in future clinical trials. At the worst, this means that regulatory approval may be sought on an endpoint that does not confirm clinical benefit for patients. The work describing the investigation of this method has also been published in Pharmaceutical Statistics (Dimier and Todd, 2017).

Examination of the information theory method identified a number of concerns with previous investigation of this approach which required resolution, including estimation of model parameters and correction of apparent errors in programming code. Results of the subsequent simulation study required closer examination to understand the impact of these changes on the performance of the approach. This research has demonstrated that estimation of individual-level surrogacy is consistently lower than expected, with extremely high variability. Whilst truly poor surrogates could be reliably identified, surrogate endpoints with medium to strong strengths of association were difficult to identify. This suggests that while the information theory approach is easy to calculate and simple to understand, it may be rare that use of the measure would allow a firm conclusion that a surrogate endpoint is suitable for use. In practice, therefore, the approach could lead to the loss of truly strong surrogate endpoints, or indeed erroneous conclusions that mediocre surrogates are good enough to be used in confirmatory studies.

As a result of these findings, a new method for the evaluation of individual-level surrogacy is proposed. Taking an approach originally intended for the purpose of building prognostic models, an extension has been developed to adequately account for the specifics of surrogacy evaluation and provide a measure that can capture the value of a proposed surrogate endpoint. This development was conducted in such a way as to maintain the overall benefits of the concept, most notably the ease of computation and interpretation. An extensive investigation of this newly developed approach to the evaluation of individual-level surrogacy demonstrates that the new methodology is able to adequately and reliably evaluate individual-level surrogacy across a range of settings, and particularly those of primary interest in this research.

A comparison of the three surrogacy approaches highlights a number of valuable findings. When considering time-to-progression as the potential surrogate endpoint, the two-stage meta-analytic copula method performs well, with low bias compared to the true underlying strength of surrogacy, and reasonably low variability. In contrast, the information theory approach performed poorly, with estimated individual-level surrogacy much

lower than the true value and with large variability, making it very difficult to draw reliable conclusions. The results of the new measure $TG_{STD,Z}(t)$ lie somewhere between these two, with estimates that are higher than those of the information theory method, and with lower variability, but lower than the results of the two-stage meta-analytic copula method, and with higher variability. Further investigation of misspecified models demonstrated that performance of the two-stage meta-analytic copula method deteriorates when moving from the ideal scenario, leading to estimates of individual-level surrogacy being comparable between this and the newly proposed method, $TG_{STD,Z}(t)$.

In contrast, when considering PFS as the potential surrogate endpoint, the performance of the two-stage meta-analytic copula method deteriorates substantially, making it difficult to recommend this measure for use in general. The information theory approach is equally limited, with under-estimated strength of surrogacy and continued high variability. This setting therefore has the greater need for improved methodology, particularly since progression-free survival is used more commonly in oncology clinical trials and is well understood and recognised as a measure of clinical benefit. The performance of the new measure $TG_{STD,Z}(t)$ has been thoroughly investigation and appears to be strong for this setting, suggesting that the strength of individual-level surrogacy can be reliably detected.

7.2 Discussion and Recommendations

When considering future evaluation of surrogate endpoints, this research provides valuable insights into issues that are previously unexplored, highlighting a number of findings and allowing key recommendations to be made. Most notably, the two-stage meta-analytic copula method, the most commonly used approach for the evaluation of time-to-event surrogate and true endpoints, has been shown to be inappropriate for universal use when considering progression-free survival as a surrogate for overall survival. Whilst the measure performs reasonably well under certain circumstances, the impact of censoring and of incorrectly specified dependence structures leads to results that can cause false conclusions

of reliable surrogacy. This finding is particularly relevant, since there have been a number of applications of the approach to this setting, for a variety of cancers (Chibaudel et al., 2011; Foster et al., 2011, 2015). Whilst Burzykowski et al. (2001) acknowledge the need for careful consideration of the choice of copula due to the potential for bias, the approach continues to be used in real-life applications with minimal discussion of the consequences. To address the problem, Alonso et al. (2017) discuss the potential to use an alternative approach that separates progression-free survival into the individual components of the endpoint (disease progression and death), but this approach has been subject to criticism, as the resulting surrogacy estimates then only reflect the individual disease states and not the overall endpoint (Ghosh et al., 2012).

This research also highlights deficiencies in the information theory approach when based on time-to-event outcomes. The measure of association used within this approach to estimate the information gain has been found to provide values that are much lower than what would be expected based on the underlying strength of surrogacy and the estimated covariate effects. The implications of this are that the approach would likely not reach values that would be high enough to conclude that a surrogate endpoint can adequately predict the true outcome for a patient. Further, the high variability in estimates when based on small sample sizes indicates that it is very difficult to achieve reliable conclusions. Despite the benefits of the information theory concept, it is difficult to conclude that the currently proposed approach can be used in practice when assessing time-to-event endpoints. It should be noted that these limitations are not relevant for the information theory approach when applied to endpoints of a different type, which is still considered to offer benefits in such settings (Alonso et al., 2017).

The measure of Total Gain developed in this research, $TG_{STD,Z}(t)$, offers an alternative approach to the evaluation of individual-level surrogacy that shares many of the benefits of the information theory approach, but is less susceptible to the drawbacks of that approach. Being computationally simple to calculate, it also offers substantial benefit over the two-stage meta-analytic copula method, which is hampered by complex numerical processes

and is subject to convergence issues. Results have demonstrated that $TG_{STD,Z}(t)$ provides reliable estimates of the underlying strength of association between surrogate and true endpoints through quantification of the ability of the surrogate to predict the true outcome, after adjusting for treatment effects on both endpoints. The measure has been shown to be largely unaffected by censoring, with only minor changes based on the dependence structure of the datasets being analysed.

Based on the totality of findings from this research, the following recommendations are considered to be appropriate. Firstly, when considering a surrogate endpoint that does not capture information directly relating to the true clinical endpoint, such as time-to-progression, it is recommended that both the two-stage meta-analytic copula method and the new measure, $TG_{STD,Z}(t)$ are used to evaluate individual-level surrogacy. The rationale for this recommendation is based on the strong performance of the former method under correctly specified models, and the finding that the $TG_{STD,Z}(t)$ measure may provide slight under-estimation of the true underlying surrogacy when assessing time-to-progression as a surrogate endpoint. Whilst the two-stage meta-analytic copula method was found to be impacted slightly by the change in dependence structure, this only affected the ability of the method to identify truly strong surrogate endpoints when there was a high proportion of patients censored. In addition, alternative copula models could be used in the estimation process if these were considered to better reflect the observed data dependence structure. The robustness of the $TG_{STD,Z}(t)$ measure to censoring would hopefully support the findings from the two-stage meta-analytic copula method, leading to reasonably similar results and conclusions. It is not recommended that the information theory approach be used, unless the amount of data available for analysis is substantially larger than that examined herein.

When the more appropriate surrogate endpoint is one that also contains information directly related to the true endpoint, such as progression-free survival, it is recommended that the $TG_{STD,Z}(t)$ measure be used for the primary evaluation of surrogacy. The results of the simulation studies presented herein have demonstrated that $TG_{STD,Z}(t)$ is

robust to censoring, is estimated with reasonable precision, and importantly reflects and distinguishes the underlying strength of association between endpoints. The timing of the analysis, as mentioned above, is considered important, and it is recommended that this be within the middle range of the Kaplan-Meier distribution for the true endpoint, such that there are sufficient patients with events of interest, and sufficient patients remaining at risk of an event, for the difference in predicted survival probabilities to remain meaningful. Use of the two-stage meta-analytic copula method and the information theory method is not recommended for this setting.

7.3 Further Work

While this research leads to a number of original findings, and allows for recommendations for the future statistical evaluation of surrogate endpoints, there are a number of potential areas for further research. Most notably, the $TG_{STD,Z}(t)$ method developed as part of this research is focused on individual-level surrogacy, but may also offer advantages in the assessment of trial-level surrogacy. The interpretation of the measure would remain intuitive, as the ability of the treatment effect on the surrogate endpoint to predict treatment effect on the true endpoint. Whilst alternative models would need to be selected to calculate the required parameters, the concept of Total Gain would be the same. Total Gain is already developed for binary and survival settings, and the extension to alternative models would be possible. Given the promising performance of Total Gain even under the small sample sizes considered here, such a development would allow further investigation into the assessment of surrogacy from these small sample sizes, to determine whether it is possible for trial-level association to be reliably estimated.

Throughout this thesis, consideration has been for the common setting in which time-to-event endpoints are analysed using proportional hazards models or Kaplan-Meier estimation. In practice, extensions or alternatives to these models may be of interest, for example those that can handle interval censoring. In many clinical trials, disease status

is determined at fixed timepoints, rather than on a continual basis, particularly when patients are required to go to hospital for their assessment. In such settings, a change in the disease can only be detected at the time of an assessment, with the actual date of the disease progression lying in the interval between disease assessments but not being specifically known. Assuming that disease progression occurs on the date of the assessment can lead to bias in estimating model parameters (Heller, 2011), and so it would be of interest to examine how these might impact the surrogacy evaluation. All three of the surrogacy methods examined are able to incorporate interval censoring through changes in the choice of models used for analysis; via the marginal distribution functions (two-stage meta-analytic copula model), the form of the relative risk (information theory method) or the models used to predict survival status at a given time (Total Gain). Examples of alternative models that can be used for interval censoring can be found in Heller (2011).

The scenarios examined in this thesis have paid close attention to the impact of surrogate endpoints that do or do not incorporate information from the true endpoint, via use of PFS and TTP respectively. Definitions of these endpoints follow recommendations of health authorities (FDA, 2007), where PFS includes the event of death, whereas TTP treats death as a censored event. It is important to note that handling of death events in this way introduces informative censoring, meaning that the censoring tells us something about the endpoint of interest. If a patient has died without prior disease progression, it is not possible to know the actual time that disease progression occurs, and moreover it may not be possible to know whether the death was caused by undetected disease progression. When censoring is assumed at the time of death, this 'drop-out' from the dataset may not therefore be independent of the actual (unobserved) event time, which invalidates the assumptions of many survival analysis techniques, including proportional hazards regression and Kaplan-Meier estimation. A number of alternative approaches are available to explore this so-called 'competing risks' setting (Austin et al., 2016), including use of the Cox proportional hazards model to estimate 'cause-specific' hazard functions, which represent the hazard of a particular event of interest, such as disease progression.

Further, cumulative incidence functions have been proposed as potential replacements of the Kaplan-Meier function, which has been shown to over-estimate the probability of observing the event of interest (Putter et al., 2007). Whilst all three surrogacy methods are able to incorporate these alternative models, their use has not been explored in the literature, and it would be of interest to further examine these options to accommodate endpoint symmetry, something that has been shown to adversely affect performance of the two-stage meta-analytic copula model. Further examination of alternative approaches would help to establish whether differences in results between TTP and PFS, or similar endpoints, can be improved.

As previously mentioned, the investigations within this research have focused on fixed values of individual and trial-level surrogacy that are assumed constant across all clinical trials included in each individually simulated meta-analysis. Further investigation using varied strengths of surrogacy may better reflect the real-life setting, and it would be of interest to understand how such changes may impact the estimation of $TG_{STD,Z}(t)$, as well as the two-stage meta-analytic copula and information theory measures.

Each of the measures described and investigated in this research provide values of surrogacy that can range between zero and one, with no threshold for determining at what point a surrogate can be considered to be established as reliable and appropriate for future use. This has been a topic of much previous discussion, with very few examples of thresholds being pre-specified before analysis of meta-analytic datasets. One example where criteria for success were pre-specified is a recent examination of response rate as a surrogate for progression-free survival (Shi et al., 2017), with consideration of thresholds for both point estimates and lower confidence limits for surrogacy estimates. It is considered likely that thresholds will be established over time, based on increased experience as well as input from regulatory authorities. Each individual disease setting may be subject to separate thresholds, depending on the severity of disease and intent of treatment. For fatal diseases where the intent is to extend life, or where there are very few treatment options available, it could be possible that more flexible thresholds are considered appropriate. In

contrast, where the intent is to improve symptoms, or where there are already a number of effective treatment options available, establishing surrogate endpoints for future use may require higher confidence and reliability in results.

Finally, it would be worthwhile for future research to explore alternative data generation methods that can accurately reflect the true underlying strength of association between endpoints without the use of a copula model. Whilst the use of lognormally distributed data was used herein to explore whether this had any impact on the performance of the three measures investigated, it would be beneficial for further work to be undertaken to determine whether more suitable approaches were possible.

The availability of appropriate statistical methodology for evaluating surrogate endpoints is a key contributing factor to their eventual use. Such methods must be designed to provide reliable and accurate conclusions, but also to be easily understandable to non-statisticians. However, the statistical methodology selected for use is also only a small part of the overall picture. Engagement between statisticians, clinicians and regulatory authorities is critical to ensure that both the statistical and non-statistical concepts of each individual surrogacy evaluation are well understood and thoroughly considered, and to further progress the use of surrogate endpoints in practice. This includes collaboration between different companies within the pharmaceutical industry, with academic and healthcare organisations, and with regulatory authorities, to establish standard practices for surrogate endpoint evaluation in a consistent and recognised framework.

In conclusion, the research presented in this thesis has provided a number of original contributions to the literature on the statistical evaluation of surrogate endpoints. The development of a new methodology to evaluate individual-level surrogacy provides an alternative option to researchers, an option that has been shown to perform well even in the setting of small sample sizes. This addition to the range of available surrogate endpoint evaluation methodologies provides a reliable approach that can be used to evaluate individual-level surrogacy in oncology settings and beyond, providing further benefit to clinical and statistical researchers attempting to tackle this important topic.

Bibliography

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

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

A. Alonso, H. Geys, G. Molenberghs, M.G. Kenward, and T. Vangeneugden. Validation of surrogate markers in multiple randomized clinical trials with repeated measurements. *Biometrical Journal*, 45:931–945, 2003.

A. Alonso, G. Molenberghs, H. Geys, M. Buyse, and T. Vangeneugden. A unifying approach for surrogate marker validation based on Prentice's criteria. *Statistics in Medicine*, 25:205–221, 2006.

A. Alonso, W. Van der Elst, G. Molenberghs, M. Buyse, and T. Burzykowski. On the Relationship between the Causal-Inference and Meta-Analytic Paradigms for the Validation of Surrogate Endpoints. *Biometrics*, 71:15–24, 2015.

A. Alonso, T. Bigirumurame, T. Burzykowski, M. Buyse, G. Molenberghs, L. Muchene, N.J. Perualila, Z. Shkedy, and W. Van der Elst. *Applied Surrogate Endpoint Evaluation Methods with SAS and R.* Chapman and Hall/CRC, 2017.

P. C. Austin, D. S. Lee, and J. P. Fine. Introduction to the Analysis of Survival Data in the Presence of Competing Risks. *Circulation*, 133(6):601–609, 2016.

Y.J. Bang, E. Van Cutsem, A. Feyereislova, H.C. Chung, L. Shen, A. Sawaki, F. Lordick, A. Ohtsu, Y. Omuro, T. Satoh, G. Aprile, E. Kulikov, J. Hill, M. Lehle, J. Rüschoff, and Y.K. Kang. 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*, 376: 687–697, 2010.

BIG 1-98 Collaborative Group. A Comparison of Letrozole and Tamoxifen in Postmenopausal Women with Early Breast Cancer. *New England Journal of Medicine*, 353:2747–2757, 2005.

BIG 1-98 Collaborative Group. Letrozole Therapy Alone or in Sequence with Tamoxifen in Women with Breast Cancer. *New England Journal of Medicine*, 361:766–776, 2009.

E. Bura and J.L. Gastwirth. The Binary Regression Quantile Plot: Assessing the Importance of Predictors in Binary Regression Visually. *Biometrical Journal*, 43:5–21, 2001.

T. Burzykowski. Validation of Surrogate Endpoints From Multiple Randomized Clinical Trials With a Failure Time True Endpoint. Unpublished Ph.D. dissertation, Limburgs Universitair Centrum, 2001.

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

T. Burzykowski, G. Molenberghs, M. Buyse, H. Geys, and D. Renard. Validation of surrogate endpoints in multiple randomized clinical trials with failure time endpoints. *Applied Statistics*, 50:405–422, 2001.

T. Burzykowski, G Molenberghs, and M. Buyse. The validation of surrogate endpoints by using data from randomized clinical trials: a case study in advanced colorectal cancer. *Journal of the Royal Statistical Society (Series A)*, 167:103–124, 2004.

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

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

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

Cardiac Arrhythmia Suppression Trial II Investigators. Effect of the Antiarrhythmic Agent Moricizine on Survival after Myocardial Infarction. *New England Journal of Medicine*, 327:227–233, 1992.

C. Chen, H. Wang, and S. Snapinn. Proportion of treatment effect (PTE) explained by a surrogate marker. *Statistics in Medicine*, 22:3449–3459, 2003.

B. Chibaudel, F. Bonnetain, Q. Shi, M. Buyse, C. Tournigand, D.J. Sargent, C.J. Allegra, R.M. Goldberg, and A. de Gramont. Alternative End Points to Evaluate a Therapeutic Strategy in Advanced Colorectal Cancer: Evaluation of Progression-Free Survival, Duration of Disease Control, and Time to Failure of Strategy - an Aide et Recherche en Cancrologie Digestive Group Study. *Journal of Clinical Oncology*, 29:4199–4204, 2011.

B. Choodari-Oskooei, P. Royston, and M.K. Parmer. A simulation study of predictive ability measures in a survival model I: Explained variation measures. *Statistics in Medicine*, 31:2627–2643, 2012a.

B. Choodari-Oskooei, P. Royston, and M.K. Parmer. A simulation study of predictive ability measures in a survival model II: explained randomness and predictive accuracy. *Statistics in Medicine*, 31:2644–2659, 2012b.

B. Choodari-Oskooei, P. Royston, and M.K.B. Parmar. The extension of total gain (TG)

statistic in survival models: properties and applications. *BMC Medical Research Methodology*, 15:50, 2015.

M. Colleoni, A. Giobbie-Hurder, M.M. Regan, B. Thürlimann, H. Mouridsen, L. Mauriac, J.F. Forbes, R. Paridaens, I. Láng, I. Smith, J. Chirgwin, T. Pienkowski, A. Wardley, K.N. Price, R.D. Gelber, A.S. Coates, and A Goldhirsch. Analyses Adjusting for Selective Crossover Show Improved Overall Survival with Adjuvant Letrozole Compared With Tamoxifen in the BIG 1-98 Study. *Journal of Clinical Oncology*, 29:1117–1124, 2011.

J. C. Cortiñas and T. Burzykowski. Simplified modeling strategies for surrogate validation with multivariate failure-time data. *Computational Statistics and Data Analysis*, 54(6): 1457–1466, 2010.

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

D.R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, 34:187–202, 1972.

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

V. De Gruttola, M. Wulfsohn, M. A. Fischl, and A. Tsiatis. Modeling the Relationship Between Survival and CD4 Lymphocytes in Patients with AIDS and AIDS-Related Complex. *Journal of Acquired Immune Deficiency Syndromes*, 6:359–365, 1993.

N. Dimier and S. Todd. 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. doi: 10.1002/pst.1812.

N. Dimier, P. Delmar, C. Ward, R. Morariu-Zamfir, G. Fingerle-Rowson, J. Bahlo, K. Fischer, B. F. Eichhorst, V. Goede, J.J.M. van Dongen, M. Ritgen, S. Böttcher, A.W.

Langerak, M. Kneba, and M. Hallek. A model for predicting effect of treatment on progression-free survival using minimal residual disease as a surrogate endpoint in chronic lymphocytic leukemia. *Blood*, 126(23):720–720, 2015. ISSN 0006-4971. URL http://www.bloodjournal.org/content/126/23/720.

EMEA. ICH Topic E8: General Conditions for Clinical Trials. Note for guidance on general considerations for clinical trials. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002877.pdf, 1998. [Online; accessed 14-May-2017].

EMEA. The European Agency for the Evaluation of Medicinal Products (EMEA). Note for guidance on evaluation of anticancer medicinal products in man. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2013/01/WC500137128.pdf, 2013. [Online; accessed 14-May-2017].

H. Ensor, R.J. Lee, C. Sudlow, and C.J. Weir. Statistical approaches for evaluating surrogate outcomes in clinical trials: A systematic review. *Journal of Biopharmaceutical Statistics*, 26:859–879, 2016.

FDA. U.S. Department of Health and Human Services, Food and Drug Administration. Center for Drug Evaluation and Research. Guidance for Industry: Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics. http://www.fda.gov/downloads/Drugs/.../Guidances/ucm071590.pdf, 2007. [Online; accessed 14-May-2017].

FDA. U.S. Department of Health and Human Services, Food and Drug Administration. Center for Drug Evaluation and Research. Guidance for Industry: Expedited Programs for Serious Conditions - Drugs and Biologics. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM358301.pdf, 2014. [Online; accessed 14-May-2017].

FDA. U.S. Department of Health and Human Services, Food and Drug Administration. Center for Drug Evaluation and Research. Guidance for Industry: Human Immunodeficiency Virus-1 Infection: Developing Antiretroviral Drugs for Treatment. http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm355128.pdf, 2015. [Online; accessed 14-May-2017].

FDA. U.S. Department of Health and Human Services, Food and Drug Administration. Center for Drug Evaluation and Research Drug and Biologic Accelerated and Restricted Distribution Approvals. https://www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/HowDrugsareDevelopedandApproved/DrugandBiologicApprovalReports/NDAandBLAApprovalReports/UCM404466.pdf, 2017. [Online; accessed 03-Sep-2017].

T.R. Fleming and D.L. DeMets. Surrogate End Points in Clinical Trials: Are We Being Misled? *Annals of Internal Medicine*, 125:605–613, 1996.

T.R. Fleming and J.H. Powers. "biomarkers and Surrogate Endpoints In Clinical Trials". *Statistics in medicine*, 31(25):2973–2984, 2012.

T.R. Fleming, R.L. Prentice, M.S. Pepe, and D. Glidden. Surrogate and auxiliary endpoints in clinical trials, with potential applications in cancer and AIDS research. *Statistics in Medicine*, 13:955–968, 1994.

N.R. Foster, Y. Qi, Q. Shi, J.E. Krook, J.W. Kugler, J.R. Jett, J.R. Molina, S.E. Schild, A.A. Adjei, and S.J. Mandrekar. Tumor response and progression-free survival as potential surrogate endpoints for overall survival in extensive stage small-cell lung cancer: findings on the basis of North Central Cancer Treatment Group trials. *Cancer*, 117: 1262–1271, 2011.

N.R. Foster, L.A. Renfro, S.E. Schild, M.W. Redman, X.F. Wang, S.E. Dahlberg, K. Ding, P.A. Bradbury, S.S. Ramalingam, D.R. Gandara, T. Shibata, N. Saijo, E.E. Vokes, A.A.

Adjei, and S.J. Mandrekar. Multi-Trial evaluation of Progression-Free survival (PFS) as a Surrogate Endpoint for Overall Survival (OS) in First-Line Extensive-Stage Small Cell Lung Cancer (ES-SCLC). *Journal of Thoracic Oncology*, 10:1099–1106, 2015.

C.E. Frangakis and D.B. Rubin. Principal Stratification in Causal Inference. *Biometrics*, 58:21–29, 2002.

L.S. Freedman. Confidence intervals and statistical power of the 'validation' ratio for surrogate or intermediate endpoints. *Journal of Statistical Planning and Inference*, 96: 143–153, 2001.

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

W.A. Fuller. *Measurement Error Models*. New York: John Wiley and Sons, 1987.

D. Ghosh, J.M.G. Taylor, and D.J. Sargent. Meta-analysis for Surrogacy: Accelerated Failure Time Models and Semicompeting Risks Modeling. *Biometrics*, 68:226–232, 2012.

E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18): 2529–2545, 1999.

H.L. Greene, D.M. Roden, R.J. Katz, R.L. Woosley, D.M. Salerno, and R.W. Henthorn. The cardiac arrhythmia suppression trial: First CAST...Then CAST II. *Journal of the American College of Cardiology*, 19:894–898, 1992.

J.R. Hecht, Y. Bang, S.K. Qin, H.C. Chung, J.M. Xu, J.O. Park, K. Jeziorski, Y. Shparyk, P.M. Hoff, A. Sobrero, P. Salman, J. Li, S.A. Protsenko, Z.A. Wainberg, M. Buyse, K. Afenjar, V. Hou, A. Garcia, T. Kaneko, Y. Huang, S. Khan-Wasti, S. Santillana, M.F. Press, and D. Slamon. Lapatinib in Combination With Capecitabine Plus Oxaliplatin in Human Epidermal Growth Factor Receptor 2-Positive Advanced or Metastatic Gastric,

Esophageal, or Gastroesophageal Adenocarcinoma: TRIO-013/LOGiC - A Randomized Phase III Trial. *Journal of Clinical Oncology*, 34(5):443–451, 2016.

G. Heller. Proportional hazards regression with interval censored data using an inverse probability weight. *Lifetime Data Analysis*, 17:373–385, 2011.

Jin-Jian Hsieh. Estimation of Kendall's tau from censored data. *Computational Statistics and Data Analysis*, 54:1613–1621, 2010.

Y. Huang and P.B. Gilbert. Comparing Biomarkers as Principal Surrogate Endpoints. *Biometrics*, 67(4):1442–1451, 2011.

M.M. Joffe and T. Greene. Related Causal Frameworks for Surrogate Outcomes. *Biometrics*, 65:530–538, 2009.

E.L. Kaplan and P. Meier. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53:457–481, 1958.

M. Kendall. A new measure of rank correlation. *Biometrika*, 30:81–93, 1938.

J.T. Kent and J. O'Quigley. Measures of Dependence for Censored Survival Data. *Biometrika*, 75(3):525–534, 1988.

W.H. Kruskal. Ordinal Measures of Association. *Journal of the American Statistical Association*, 53:814–861, 1958.

S. Laporte, P. Squifflet, N. Baroux, F. Fossella, V. Georgoulias, J. Pujol, J. Douillard, S. Kudoh, J. Pignon, E. Quinaux, and M. Buyse. 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*, 3(3), 2013. ISSN 2044-6055.

Z. Li, M.P. Meredith, and M.S. Hoseyni. A method to assess the proportion of treatment effect explained by a surrogate endpoint. *Statistics in Medicine*, 20:3175–3188, 2001.

D.Y. Lin, T.R. Fleming, and V. DeGruttola. Estimating the proportion of treatment effect explained by a surrogate marker. *Statistics in Medicine*, 16:1515–1527, 1997.

I.C. Marschner, A.C. Collier, R.W. Coombs, R.T. D'Aquila, V. DeGruttola, M.A. Fischl, S.M. Hammer, D.H. Hughes, V.A. Johnson, D.A. Katzenstein, Richman D.D., L.M. Smeaton, S.A. Spector, and M.S. Saag. Use of Changes in Plasma Levels of Human Immunodeficiency Virus Type 1 RNA to Assess the Clinical Benefit of Antiretroviral Therapy. *Journal of Infectious Diseases*, 177:40–47, 1998.

A.W. Marshall and I. Olkin. Families of Multivariate Distributions. *Journal of the American Statistical Association*, 83(403):834–841, 1988.

G. Molenberghs, T. Burzykowski, A. Alonso, P. Assam, A. Tilahun, and M. Buyse. The meta-analytic framework for the evaluation of surrogate endpoints in clinical trials. *Journal of Statistical Planning and Inference*, 138:432–449, 2008.

R.G. Nelsen. *An introduction to copulas*. Springer, 1999.

J. O'Quigley. *Proportional Hazards Regression*. Springer, 2008.

J. O'Quigley and P. Flandre. Predictive capabilityof proportional hazards regression. *Proceedings of the National Academy of Sciences of the United States of America*, 91 (6):2310–2314, 1994.

J. O'Quigley and P. Flandre. Quantification of the Prentice criteria for surrogate endpoints. *Biometrics*, 62:297–300, 2006.

J. O'Quigley and R. Xu. In *Handbook of Statistics in Clinical Oncology*, chapter 19, pages 397–410. Marcel Dekker: New York, 2001.

J. O'Quigley, R.i Xu, and J. Stare. Explained randomness in proportional hazards models. *Statistics in Medicine*, 24(3):479–489, 2005.

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

C.M. Pratt and L.A. Moye. The cardiac arrhythmia suppression trial: background, interim results and implications. *Journal of the American College of Cardiology*, 65:20B–29B, 1990.

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

A. Pryseley. Marker methodology, with focus on time-to-event outcomes. Unpublished Ph.D. dissertation, Hasselt University, 2009.

A. Pryseley, A. Tilahun, A. Alonso, and G. Molenberghs. An information-theoretic approach to surrogate-marker evaluation with failure time endpoints. *Lifetime Data Analysis*, 17:195–214, 2011.

H. Putter, M. Fiocco, and R. B. Geskus. Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine*, 26(11):2389–2430, 2007.

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

L.A Renfro, Q. Shi, Y. Xue, J. Li, H. Shang, and D.J. Sargent. Center-within-trial versus trial-level evaluation of surrogate endpoints. *Computational Statistics and Data Analysis*, 78(4):1–20, 2014.

L.A Renfro, H. Shang, and D. Sargent. Impact of Copula Directional Specification on Multi-Trial Evaluation of Surrogate End Points. *Journal of Biopharmaceutical Statistics*, 25:857–877, 2015.

J.M. Robins and S. Greenland. Identifiability and exchangeability of direct and indirect effects. *International Journal of Epidemiology*, 3:143–155, 1992.

P. Royston. Explained variation for survival models. *Stata Journal*, 6(1):83–96(14), 2006. URL `http://www.stata-journal.com/article.html?article=st0098`.

P. Royston and W. Sauerbrei. A new measure of prognostic separation in survival data. *Statistics in Medicine*, 23(5):723–748, 2004.

M.l Schemper and R. Henderson. Predictive Accuracy and Explained Variation in Cox Regression. *Biometrics*, 56(1):249–255, 2000.

G. Schwarz. Estimating the Dimension of a Model. *Annals of Statistics*, 6:461–464, 1978.

C.E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 1948.

Q. Shi, L.A. Renfro, B.M. Bot, T. Burzykowski, M. Buyse, and D. Sargent. Comparative assessment of trial-level surrogacy measures for candidate time-to-event surrogate endpoints in clinical trials. *Computational Statistics and Data Analysis*, 55:2748–2757, 2011.

Q. Shi, D.J. Sargent, and L.A. Renfro. Findings from the Adjuvant Colon Cancer End Points (ACCENT) Collaborative Group: the Power of Pooled Individual Patient Data from Multiple Clinical Trials. *Current Colorectal Cancer Reports*, 12(5):251–259, 2016.

Q. Shi, C.R. Flowers, W. Hiddemann, R. Marcus, M. Herold, A. Hagenbeek, E. Kimby, H. Hochster, U. Vitolo, B.A. Peterson, E. Gyan, M. Ghielmini, T. Nielsen, S. De Bedout, T. Fu, N. Valente, N.H. Fowler, E. Hoster, M. Ladetto, F. Morschhauser, E. Zucca, G. Salles, and D.J. Sargent. 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*, 35:552–560, 2017.

P. K. Trivedi and D.M. Zimmer. Copula modeling: An introduction for practitioners. *Foundations and Trends in Econometrics*, 1:1–111, 2007.

A.A. Tsiatis. Surrogate markers in AIDS clinical trials. *Presented at the ENAR Meetings*, 50:405–422, 1996.

H.C. van Houwelingen, L.R. Arends, and T. Stijnen. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine*, 21:589–624, 2002.

T.J. VanderWeele. Simple relations between principal stratification and direct and indirect effects. *Statistics and Probability Letters*, 78:2957–2962, 2008.

C. J. Weir and R. J. Walley. Statistical evaluation of biomarkers as surrogate endpoints: a literature review. *Statistics in Medicine*, 25:183–203, 2006.

R. Xu and J. O'Quigley. A R^2 type measure of dependence for proportional hazards models. *Journal of Nonparametric Statistics*, 12:83–107, 1999.

Appendix A

Two-Stage Meta-Analytic Copula Method

Additional Results

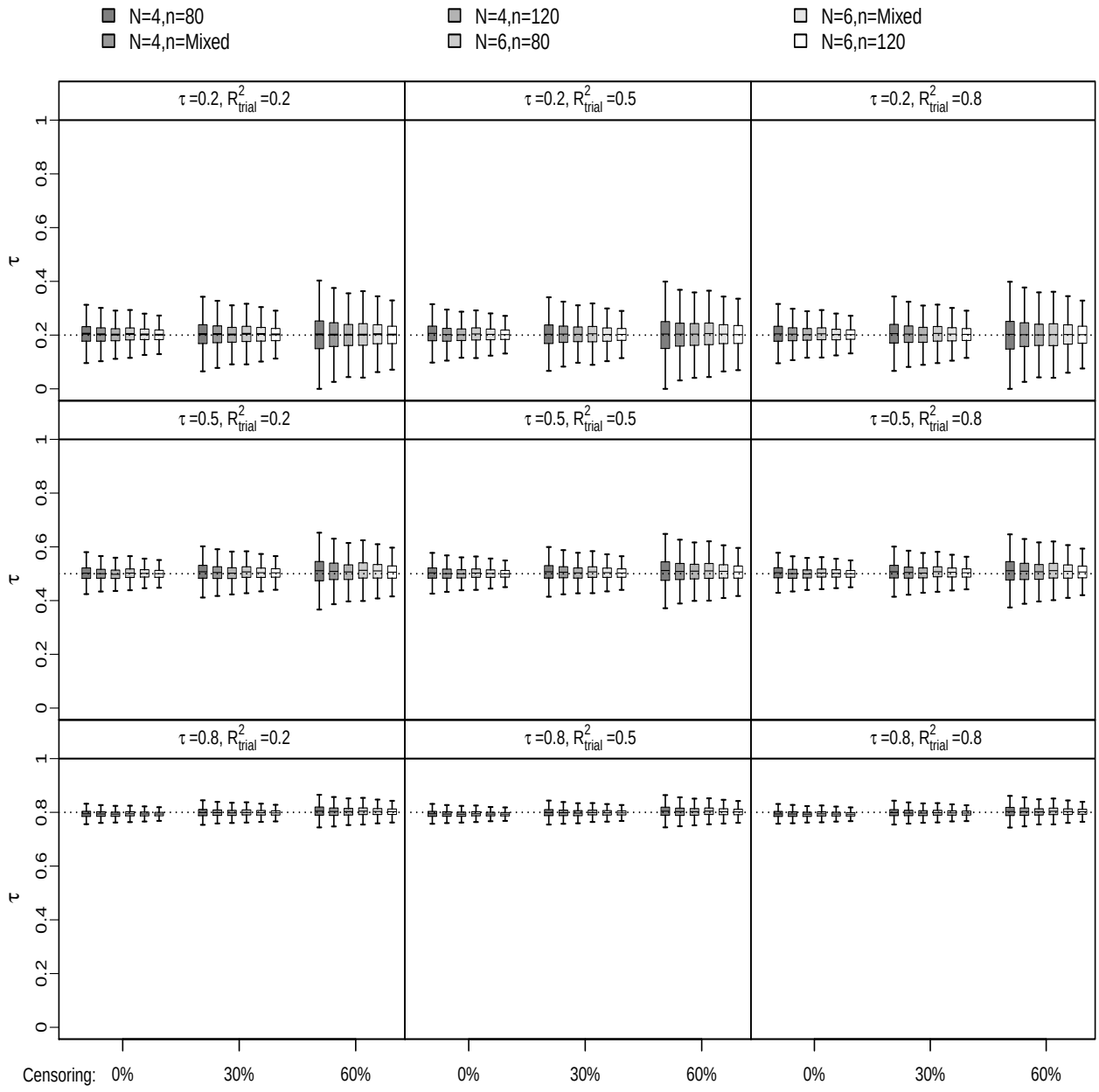

Figure A.1: Boxplots of estimates of τ: TTP, Clayton Copula Data Generation, Clayton Copula Application (wider range of treatment effects on T)

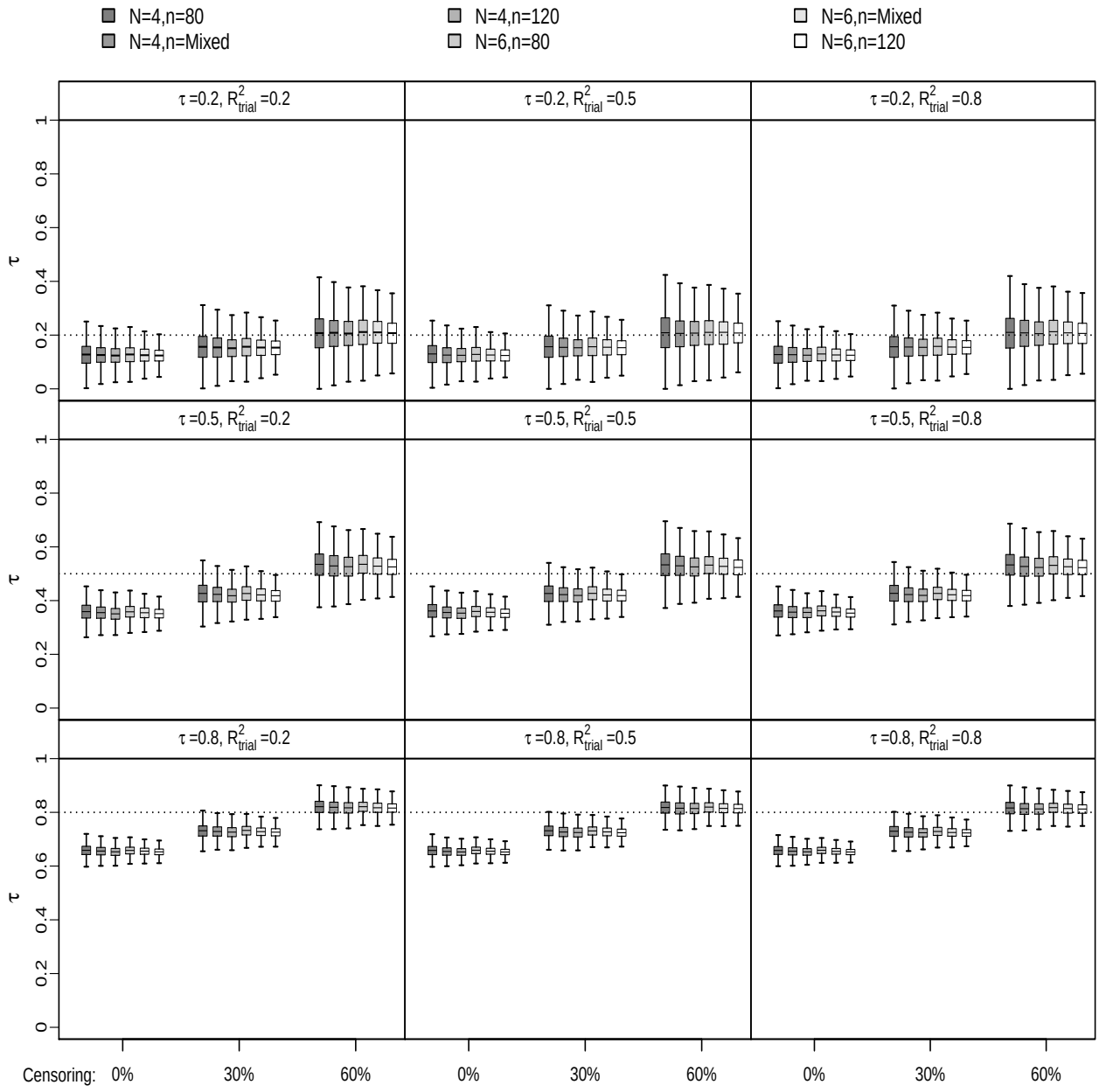

Figure A.2: Boxplots of estimates of τ: TTP, Gumbel Copula Data Generation, Clayton Copula Application (wider range of treatment effects on T)

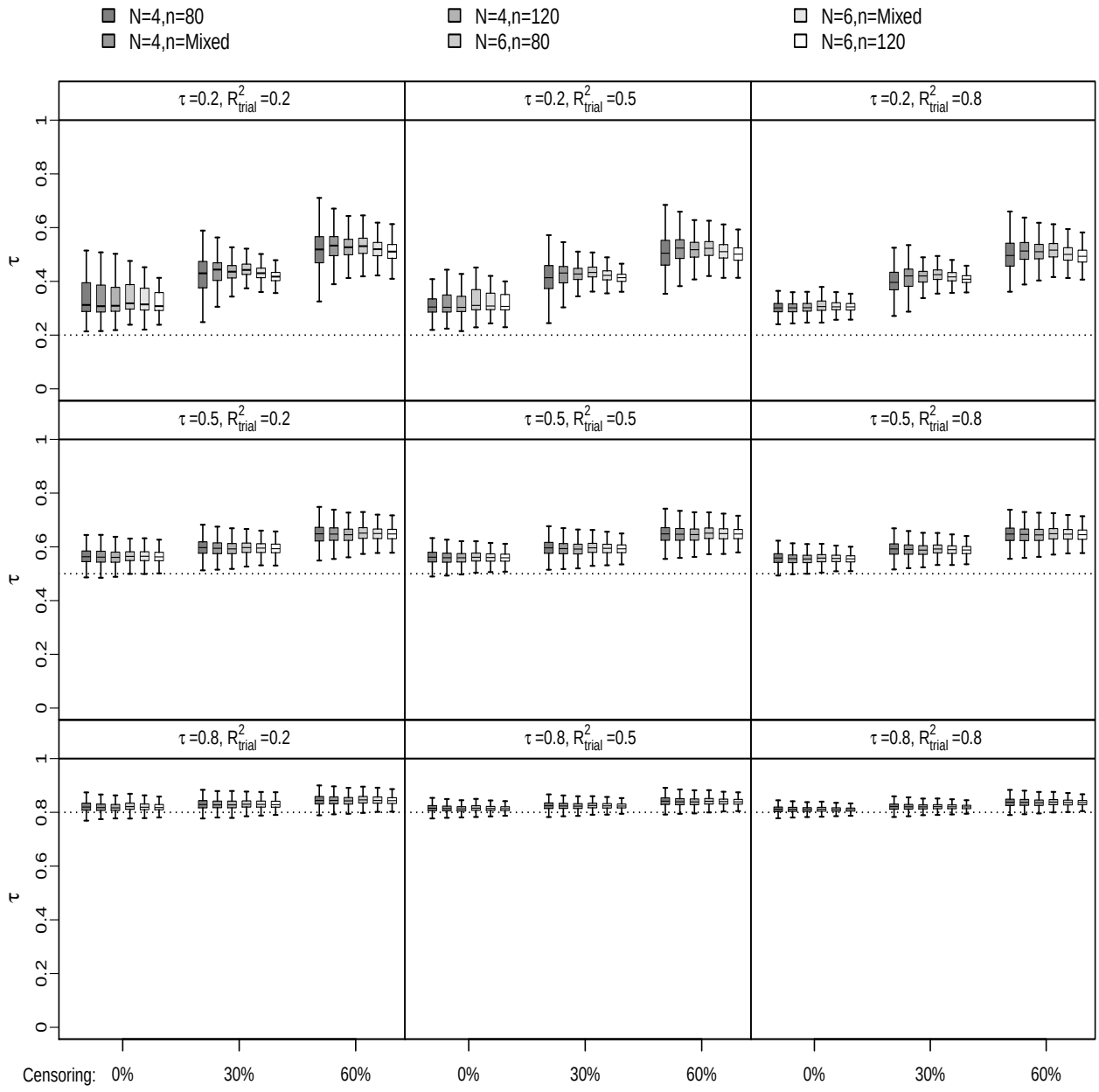

Figure A.3: Boxplots of estimates of τ: PFS, Clayton Copula Data Generation, Clayton Copula Application (wider range of treatment effects on T)

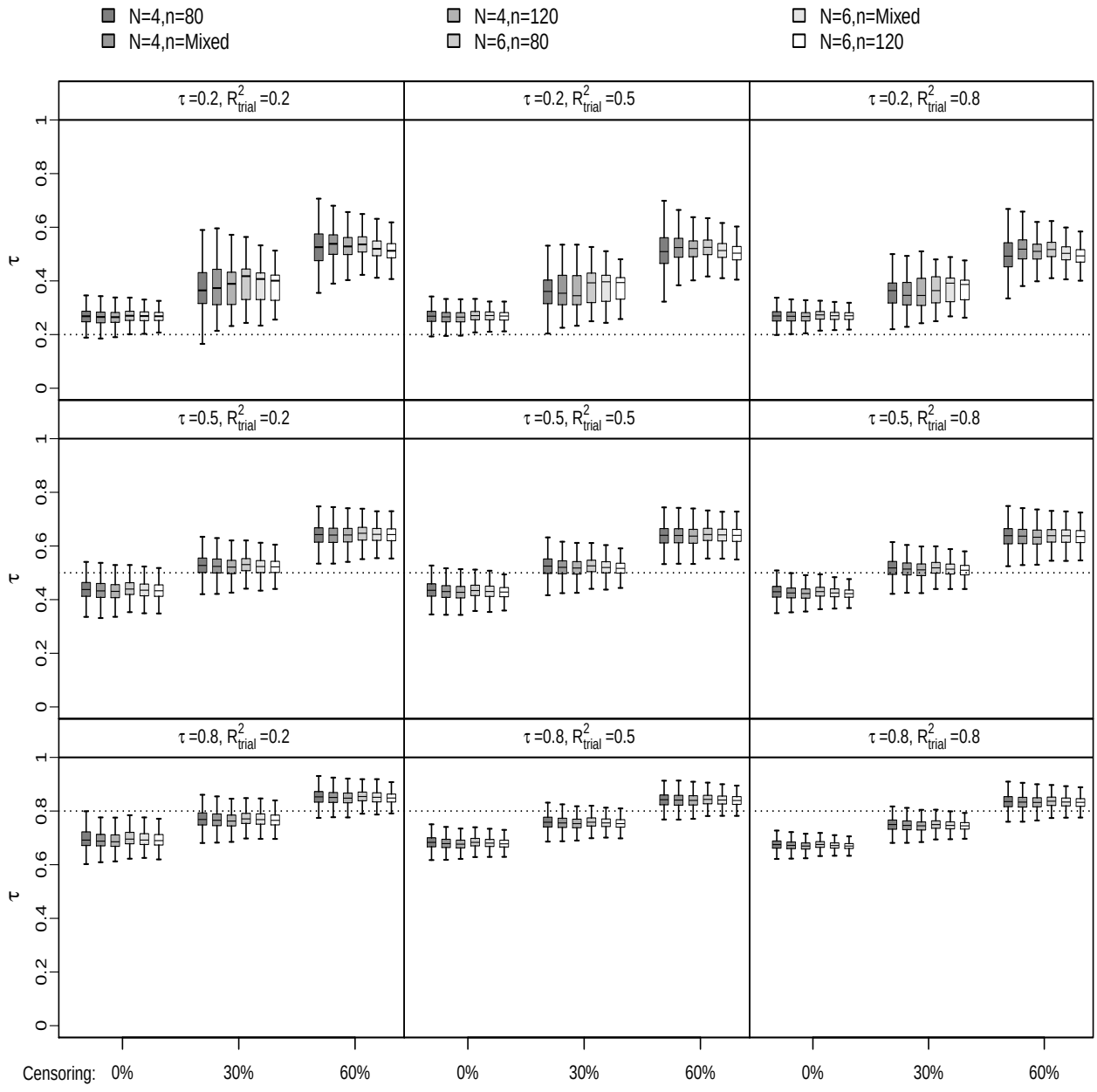

Figure A.4: Boxplots of estimates of τ: PFS, Gumbel Copula Data Generation, Clayton Copula Application (wider range of treatment effects on T)

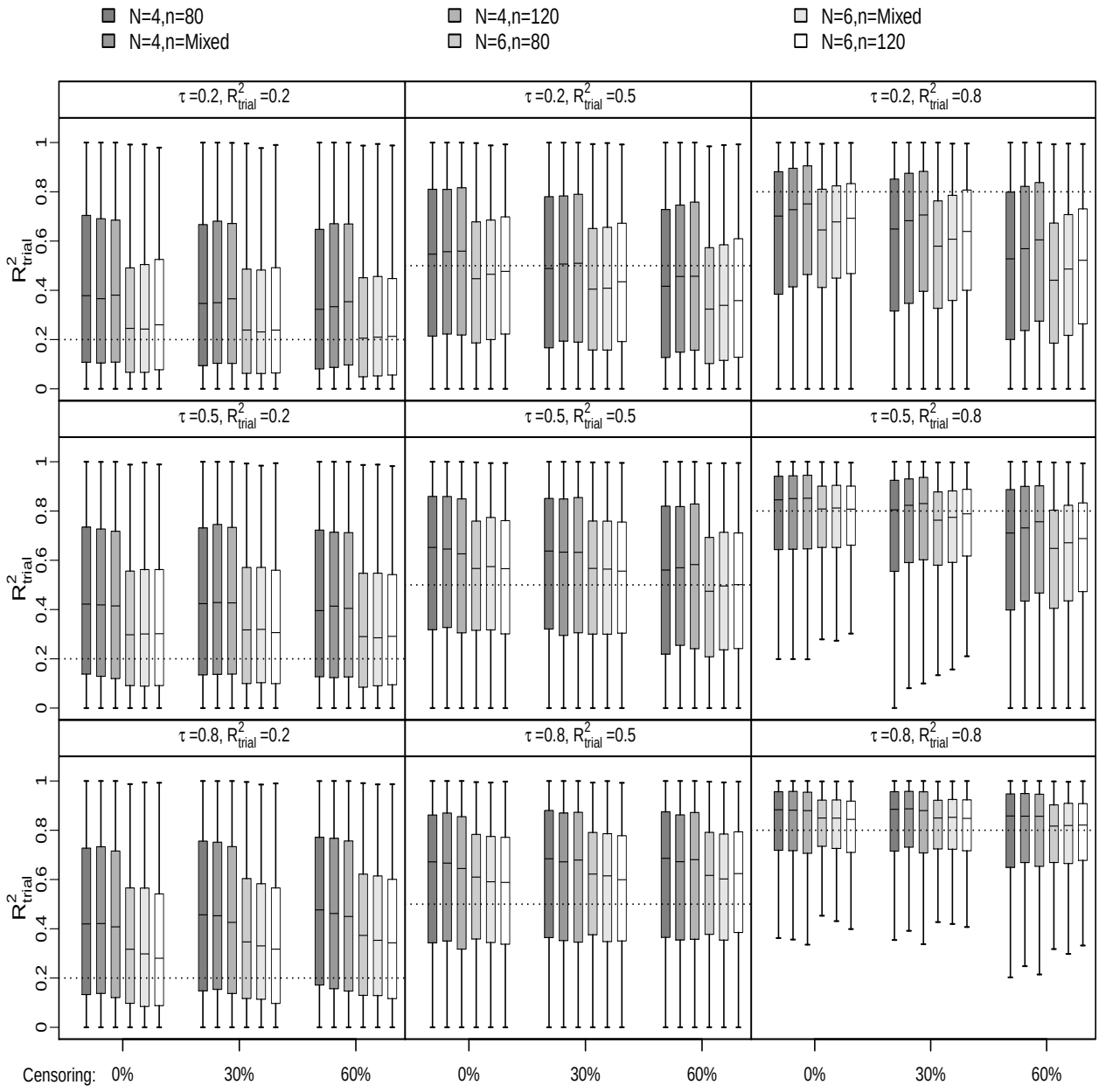

Figure A.5: Boxplots of estimates of R^2_{trial}: TTP, Clayton Copula Data Generation, Clayton Copula Application (wider range of treatment effects on T)

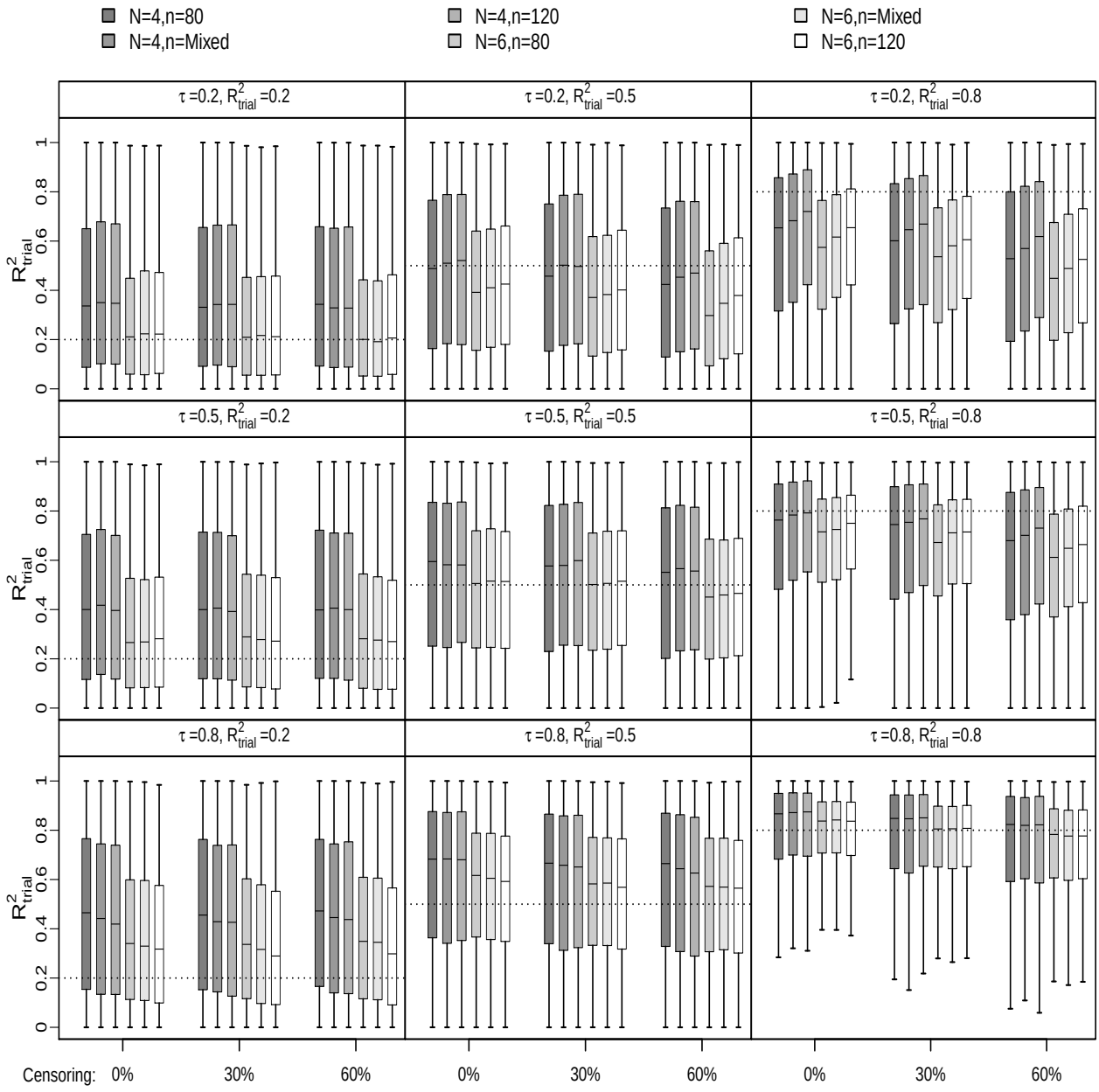

Figure A.6: Boxplots of estimates of R^2_{trial}: TTP, Gumbel Copula Data Generation, Clayton Copula Application (wider range of treatment effects on T)

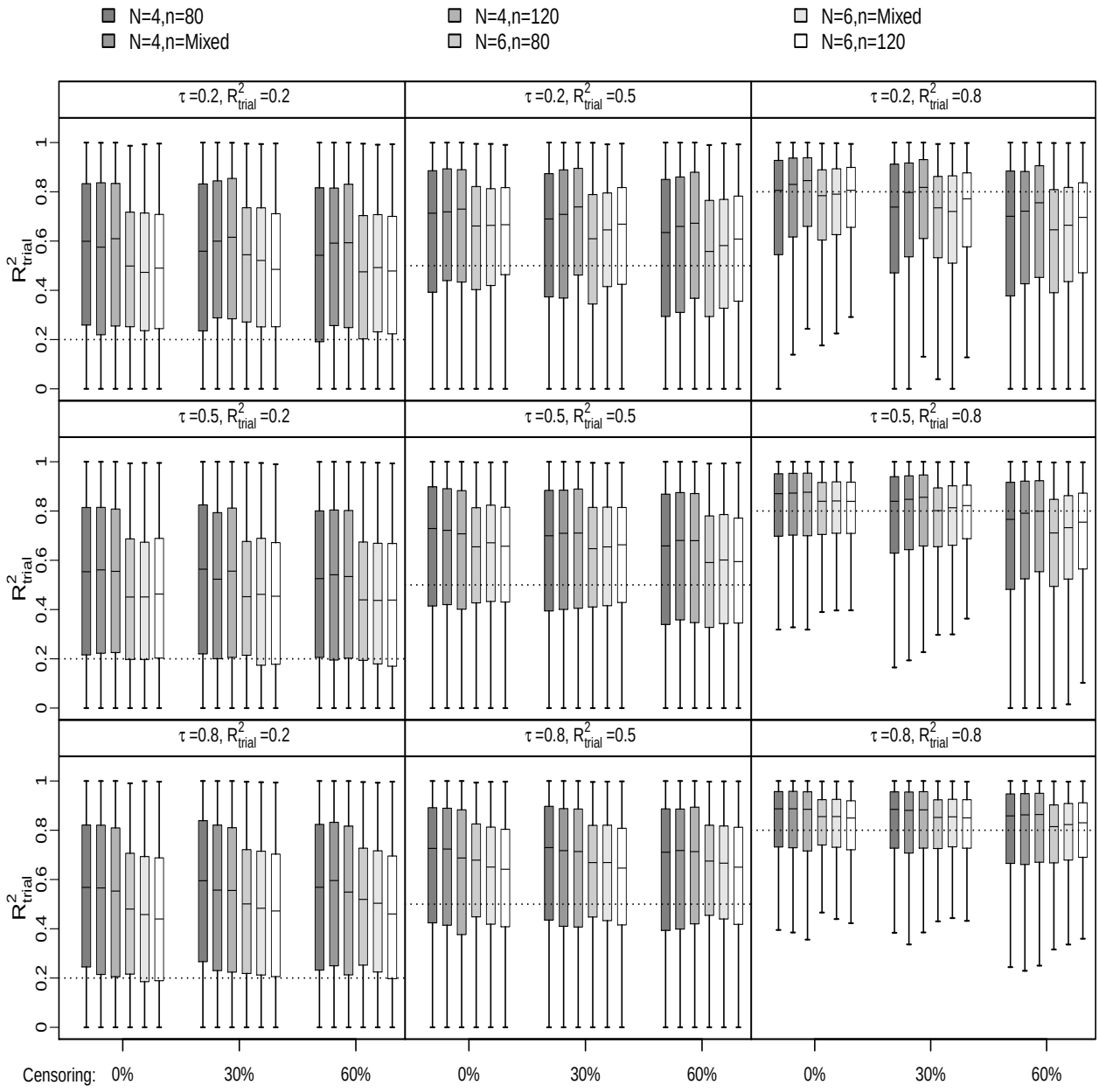

Figure A.7: Boxplots of estimates of R^2_{trial}: PFS, Clayton Copula Data Generation, Clayton Copula Application (wider range of treatment effects on T)

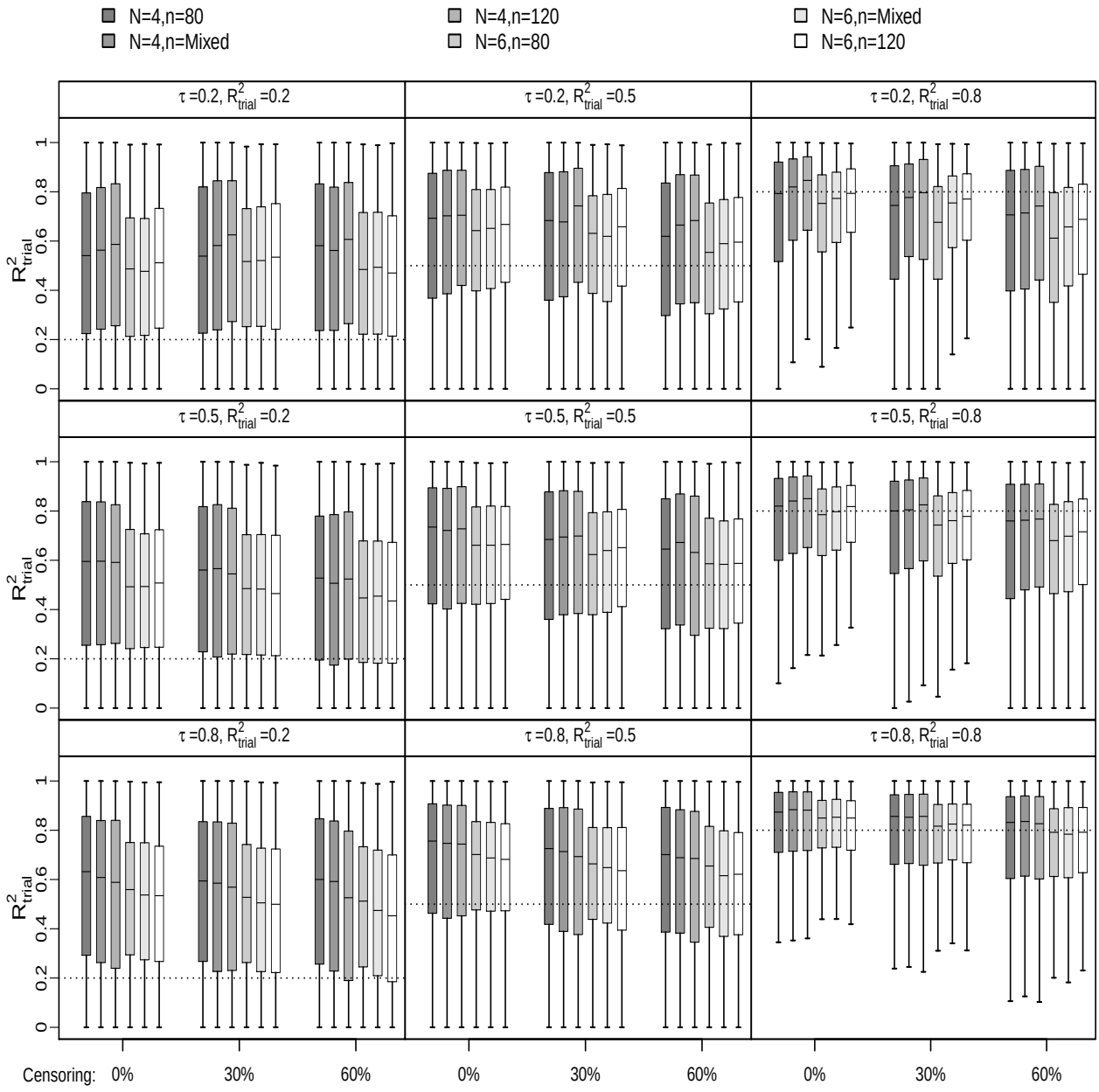

Figure A.8: Boxplots of estimates of R^2_{trial}: PFS, Gumbel Copula Data Generation, Clayton Copula Application (wider range of treatment effects on T)

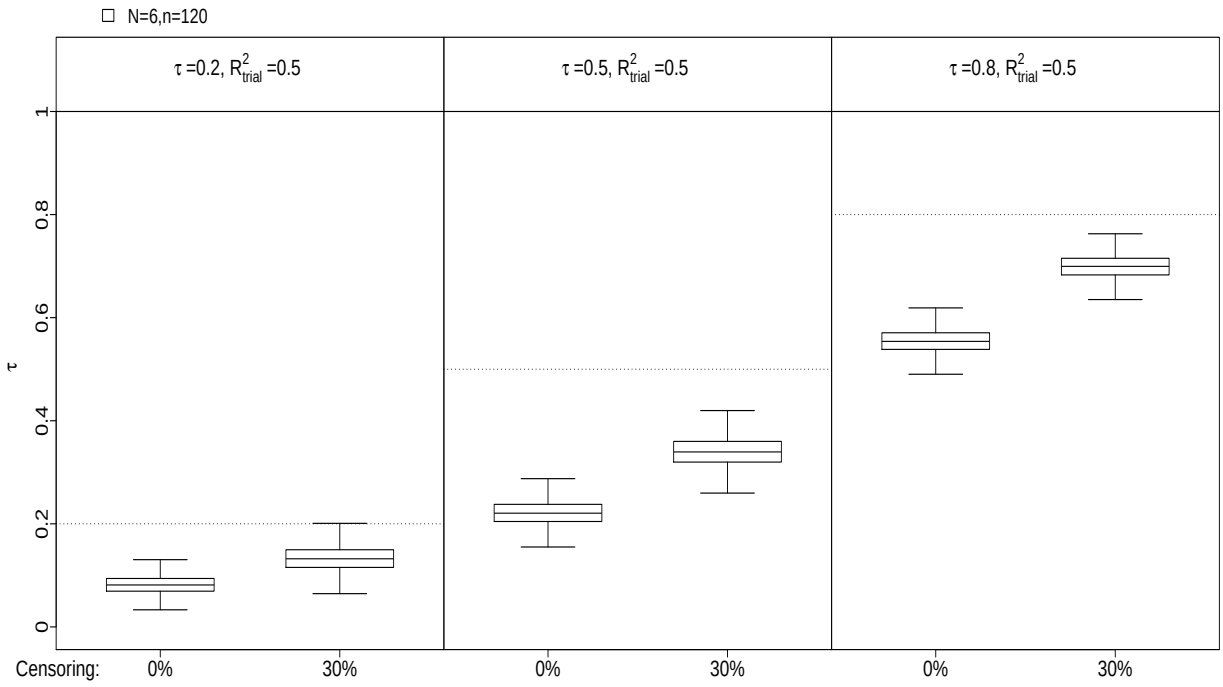

Figure A.9: Boxplots of estimates of τ: TTP, Lognormal Data Generation, Clayton Copula Application

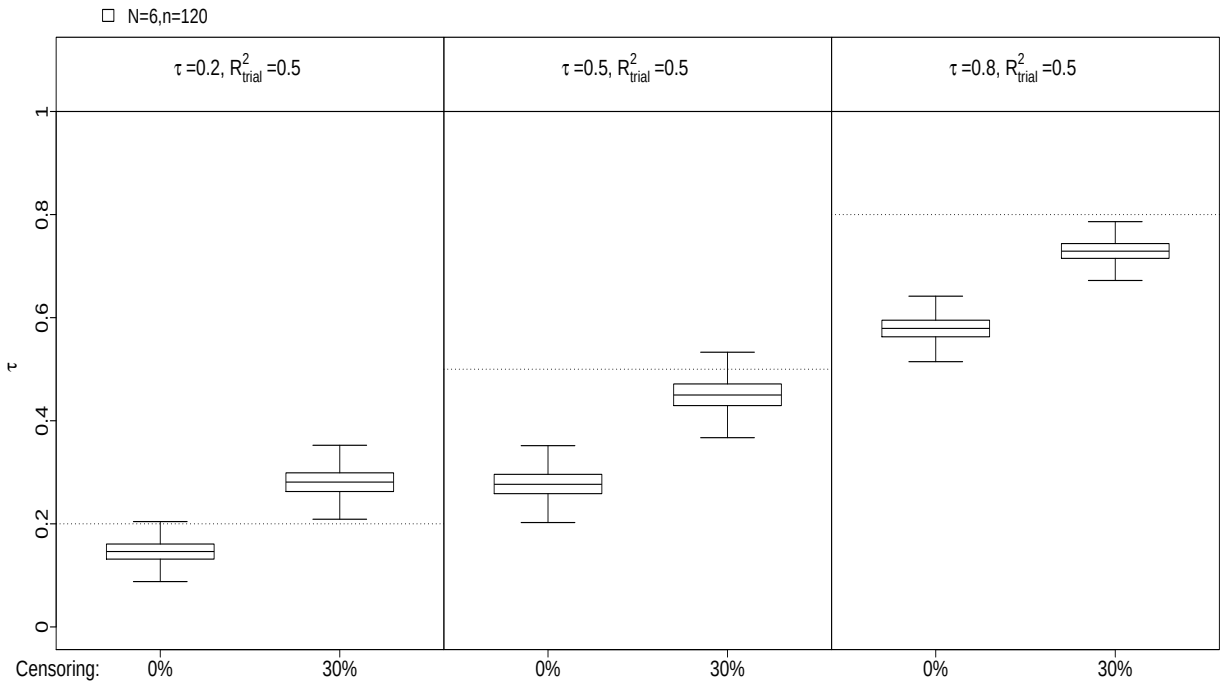

Figure A.10: Boxplots of estimates of τ: PFS, Lognormal Data Generation, Clayton Copula Application

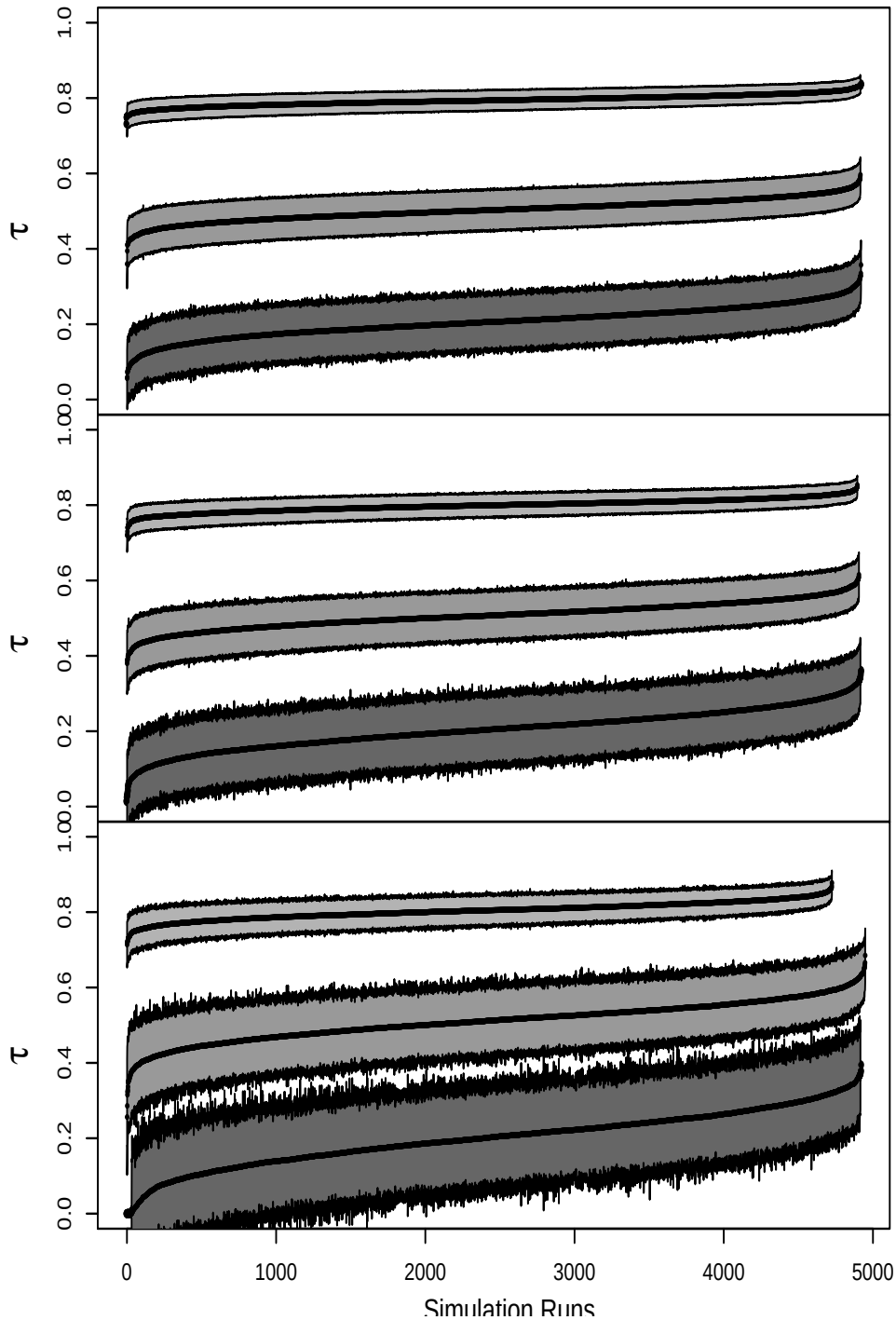

Figure A.11: Confidence Intervals for τ ($R^2_{trial} = 0.5$, $N = 4$, $n = 80$): TTP, Clayton Copula Data Generation, Clayton Copula Application (values of τ ordered from smallest to largest for easier interpretation) - 0% censoring (top row), 30% censoring (middle row), 60% censoring (bottom row)

Appendix B

Information Theory Method

Additional Results

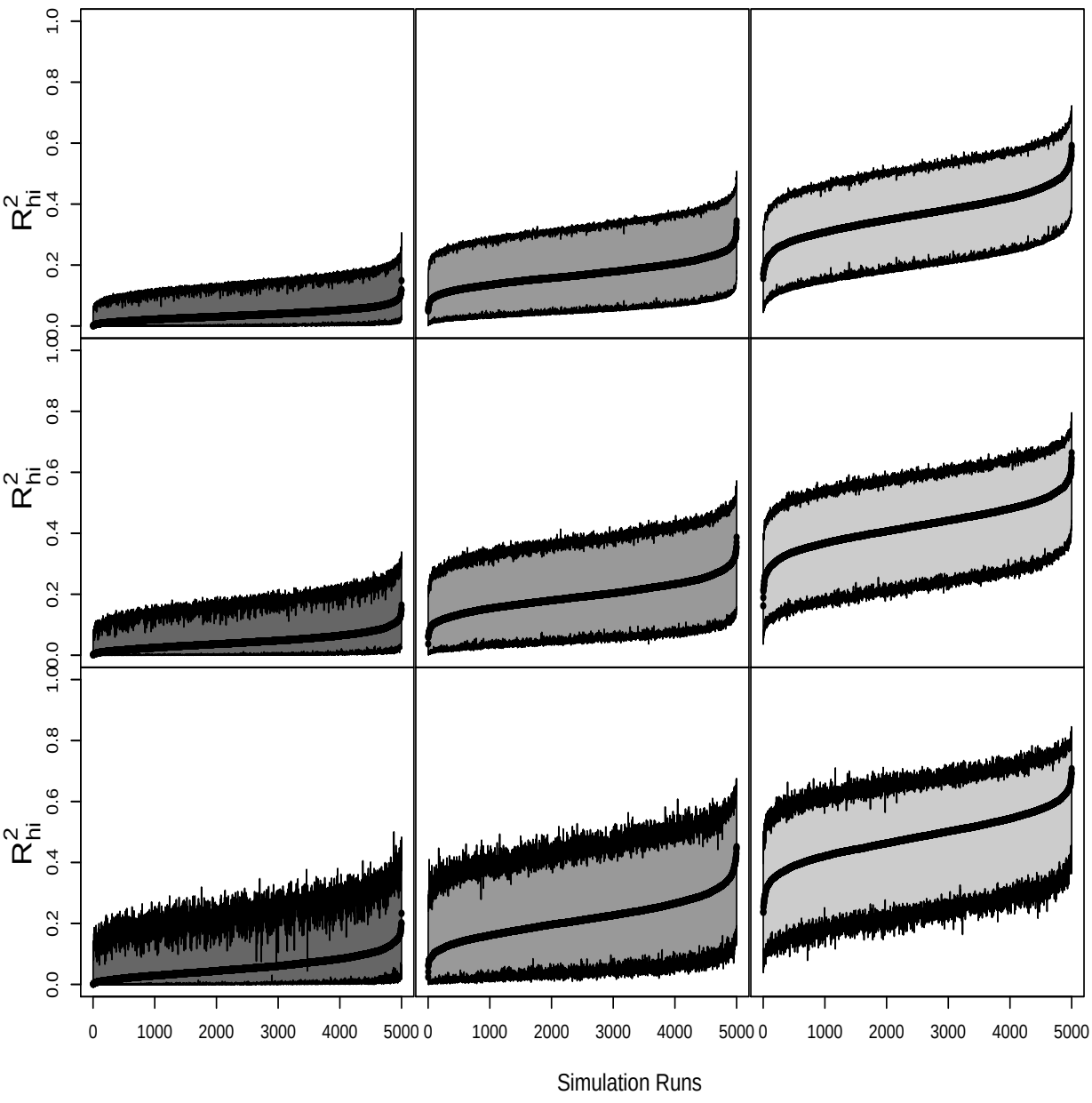

Figure B.1: Confidence Intervals for $R^2_{h,i}$ ($R^2_{trial} = 0.5$, $N = 4$, $n = 80$): TTP, Clayton Copula Data Generation, Information Theory Application (values of $R^2_{h,i}$ ordered from smallest to largest for easier interpretation) - 0% censoring (top row), 30% censoring (middle row), 60% censoring (bottom row); $\tau = 0.2$ (left column), $\tau = 0.5$ (middle column), $\tau = 0.8$ (right column)

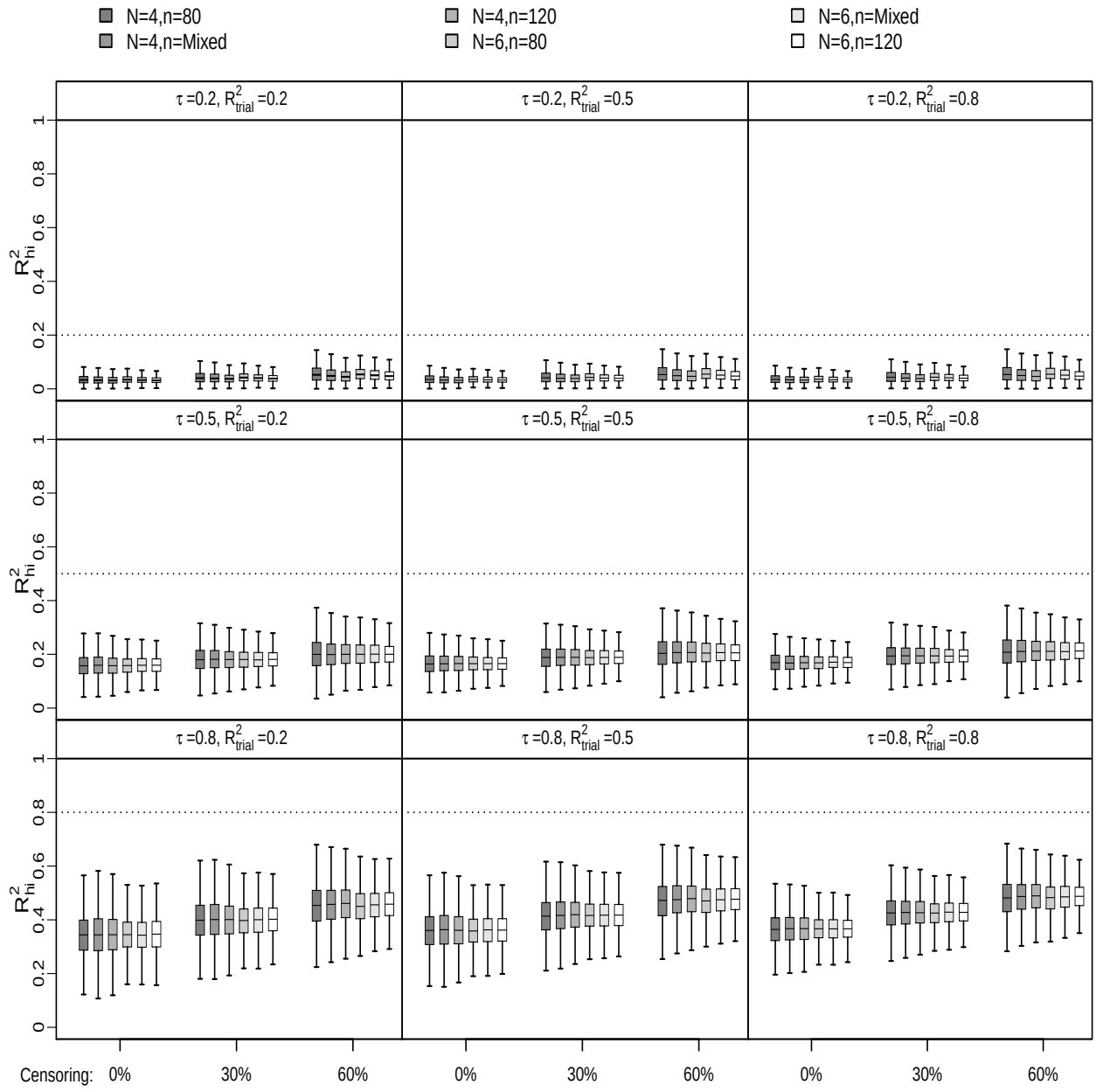

Figure B.2: Boxplots of estimates of $R^2_{h,i}$: TTP, Clayton Copula Data Generation, Information Theory Application (wider range of treatment effects on T)

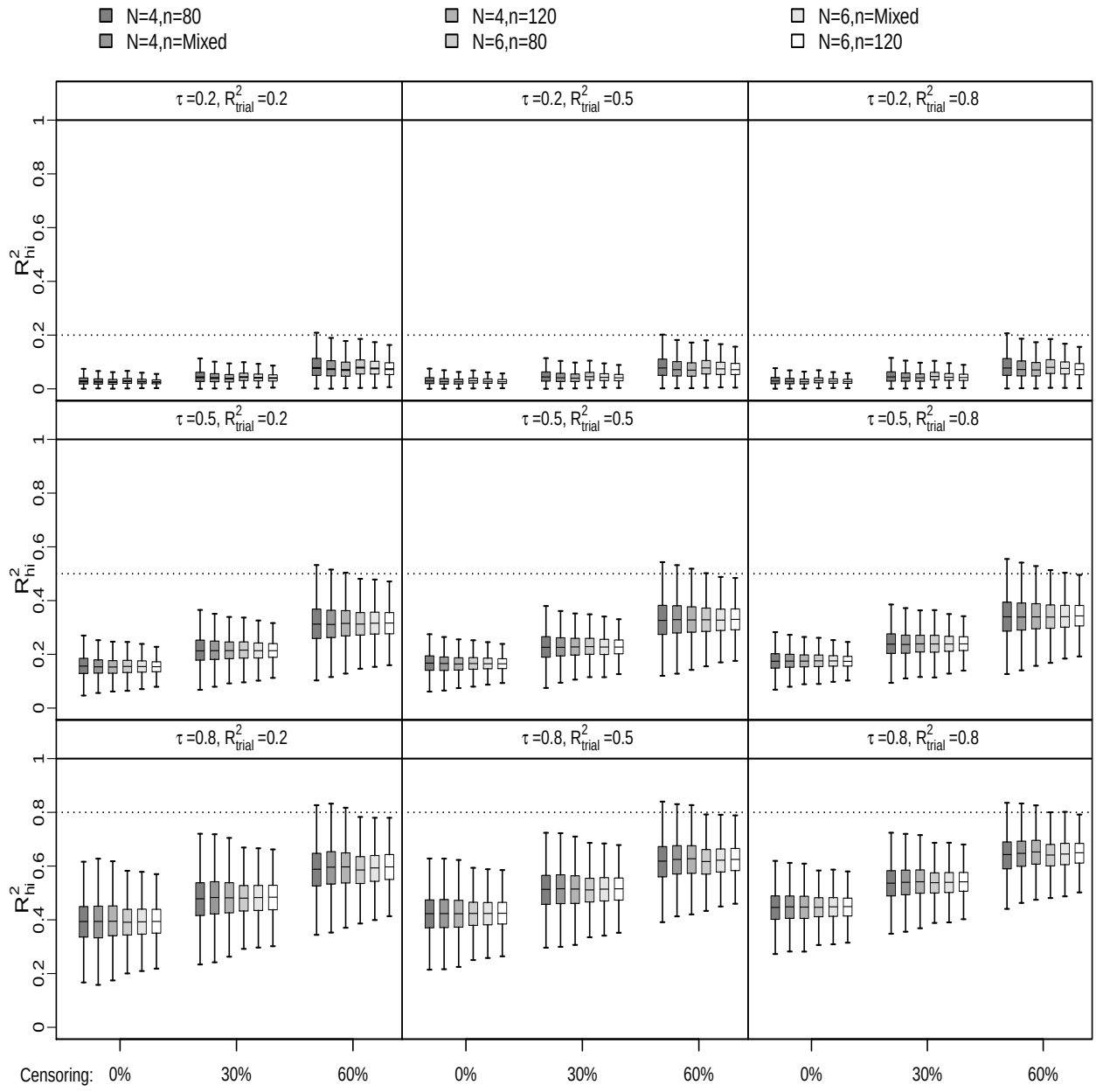

Figure B.3: Boxplots of estimates of $R^2_{h,i}$: TTP, Gumbel Copula Data Generation, Information Theory Application (wider range of treatment effects on T)

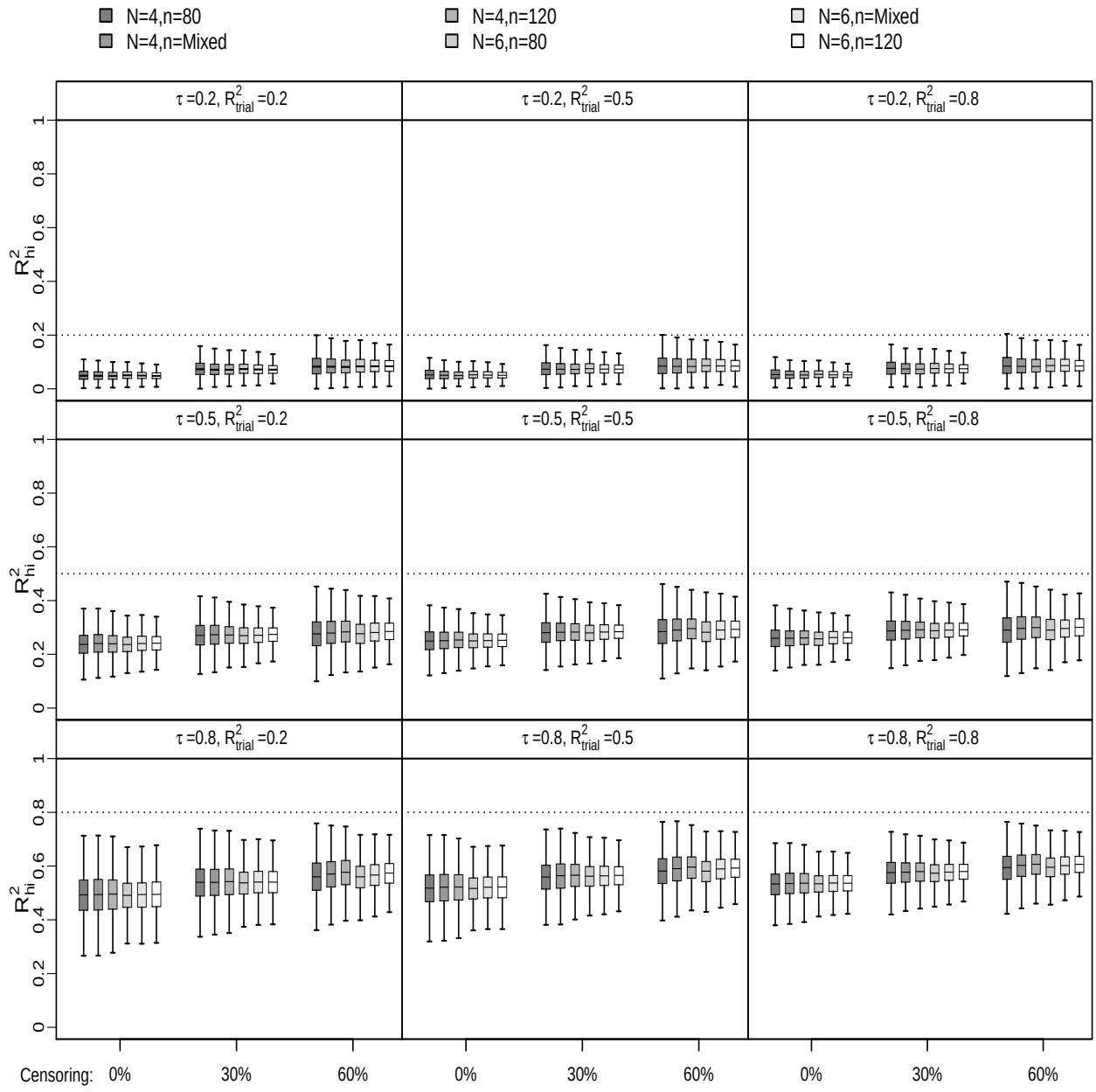

Figure B.4: Boxplots of estimates of $R^2_{h,i}$: PFS, Clayton Copula Data Generation, Information Theory Application (wider range of treatment effects on T)

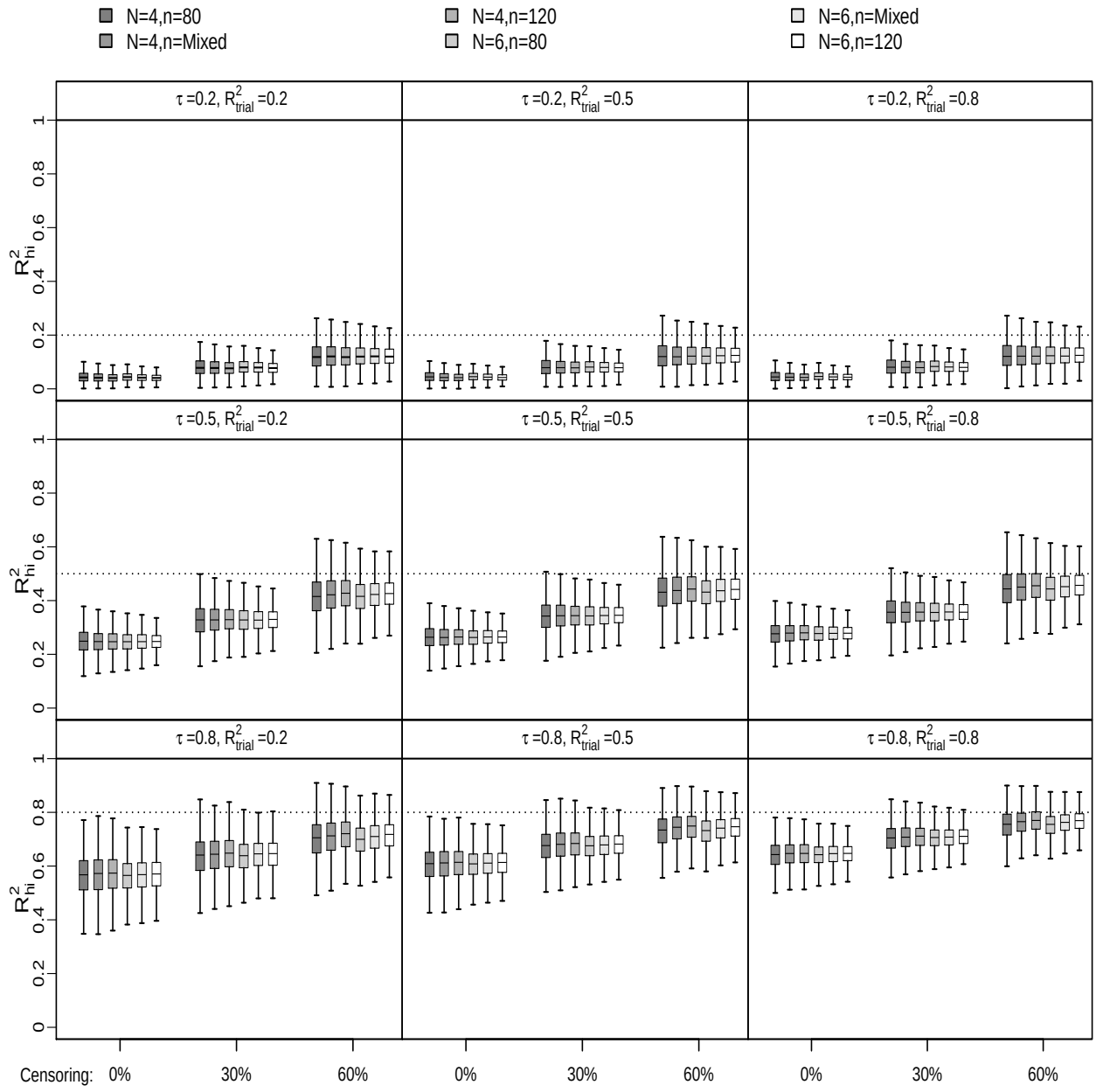

Figure B.5: Boxplots of estimates of $R^2_{h,i}$: PFS, Gumbel Copula Data Generation, Information Theory Application (wider range of treatment effects on T)

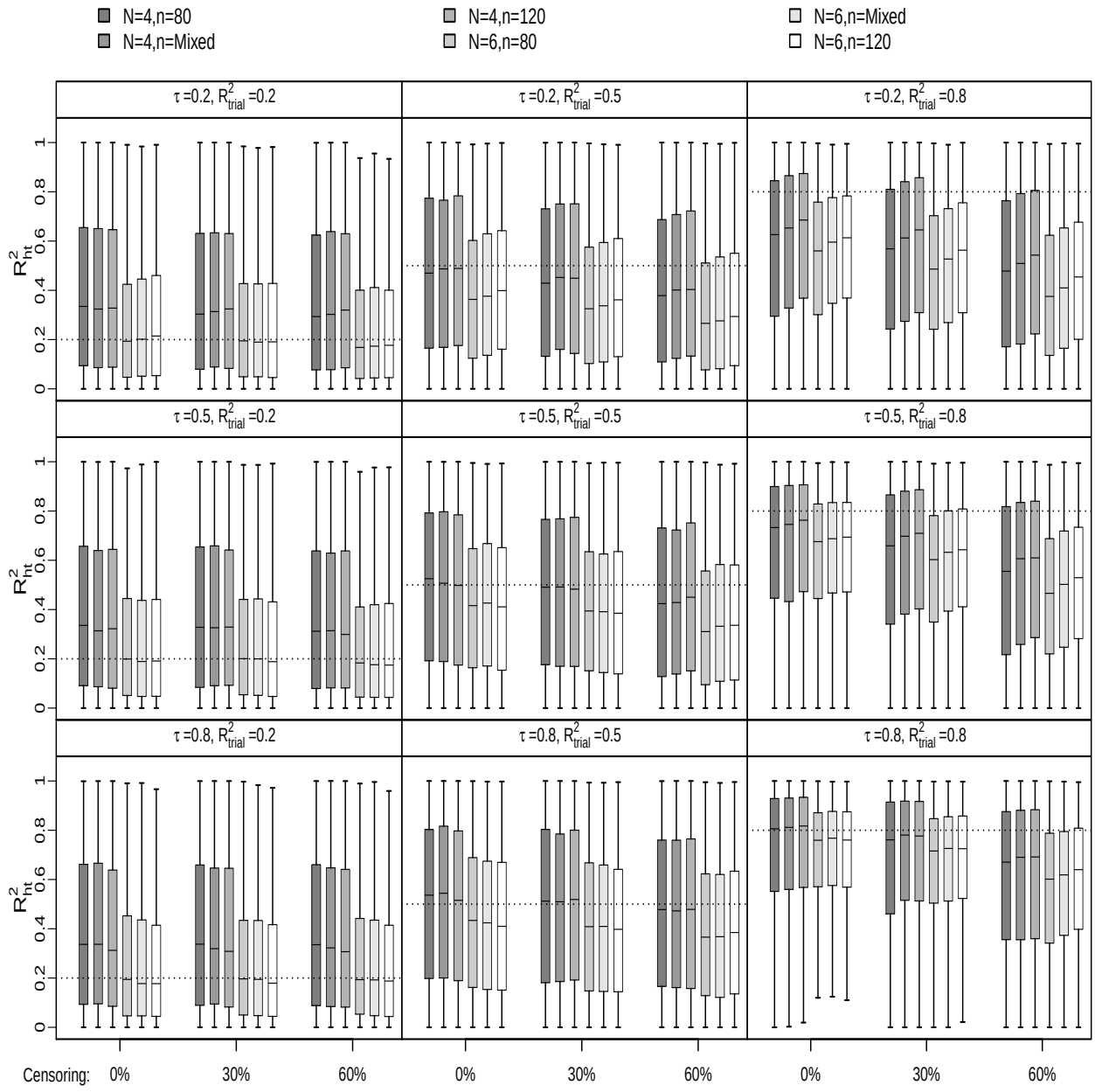

Figure B.6: Boxplots of estimates of R^2_{trial}: TTP, Clayton Copula Data Generation, Information Theory Application (wider range of treatment effects on T)

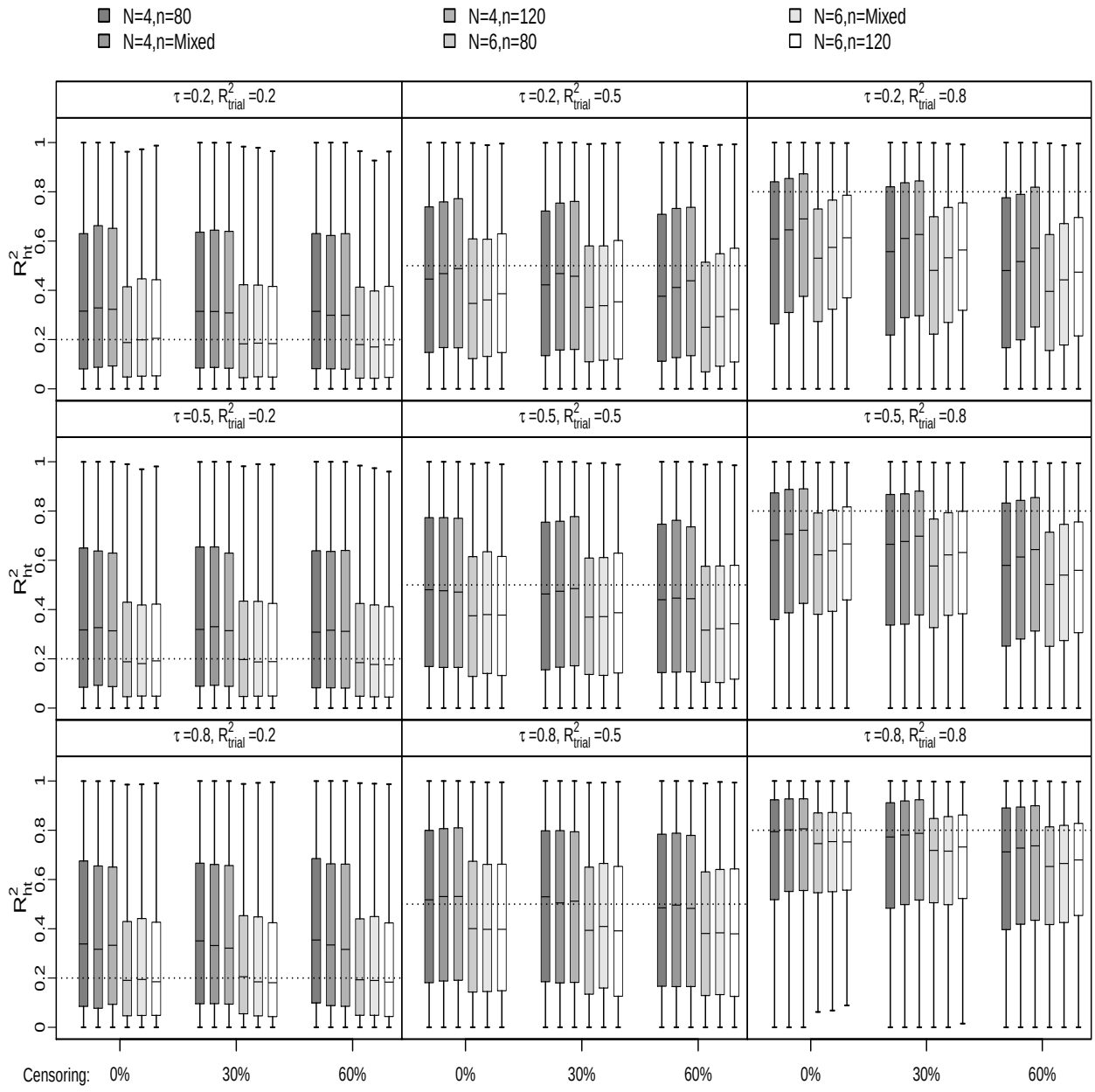

Figure B.7: Boxplots of estimates of R^2_{trial}: TTP, Gumbel Copula Data Generation, Information Theory Application (wider range of treatment effects on T)

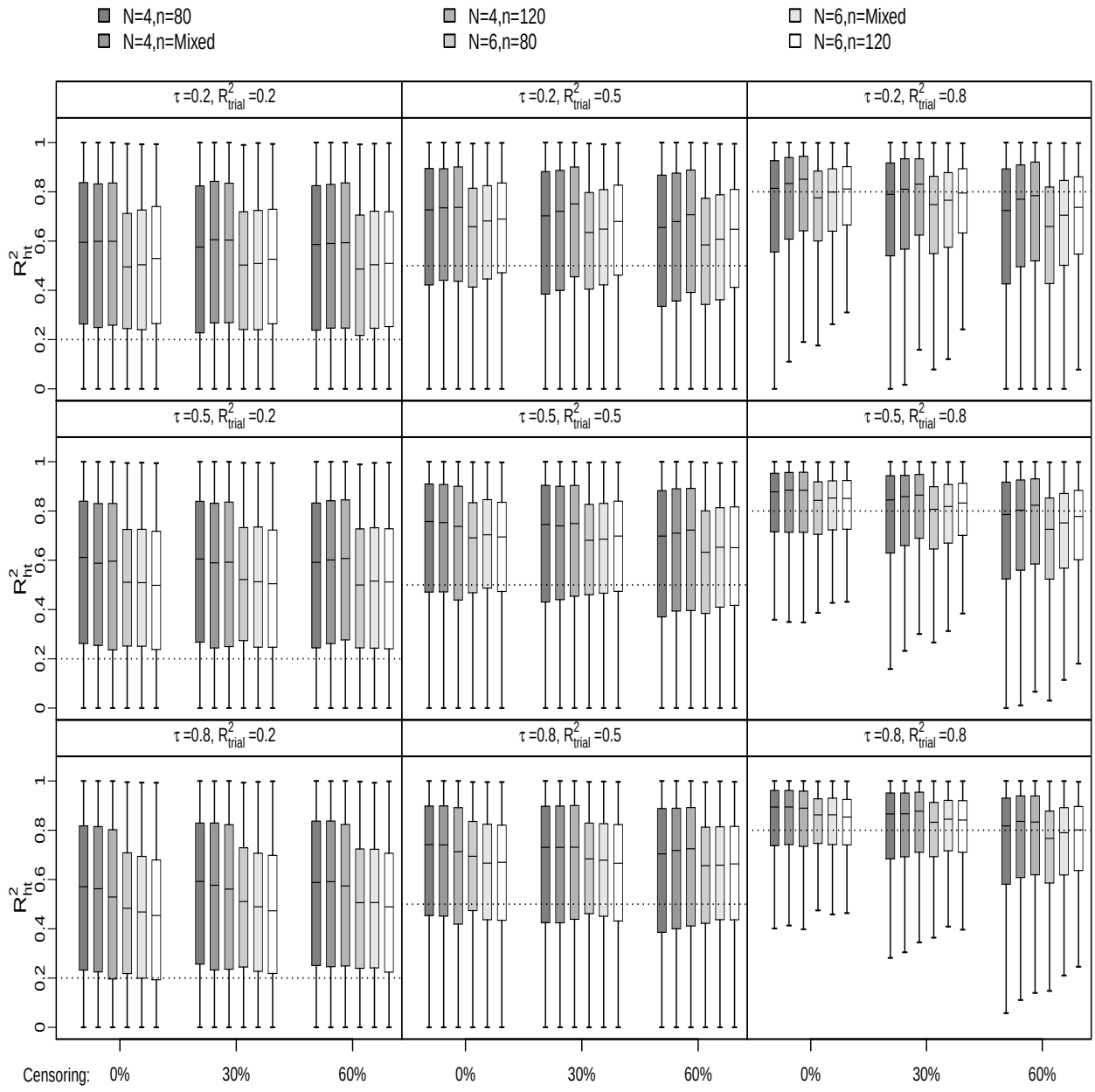

Figure B.8: Boxplots of estimates of R^2_{trial}: PFS, Clayton Copula Data Generation, Information Theory Application (wider range of treatment effects on T)

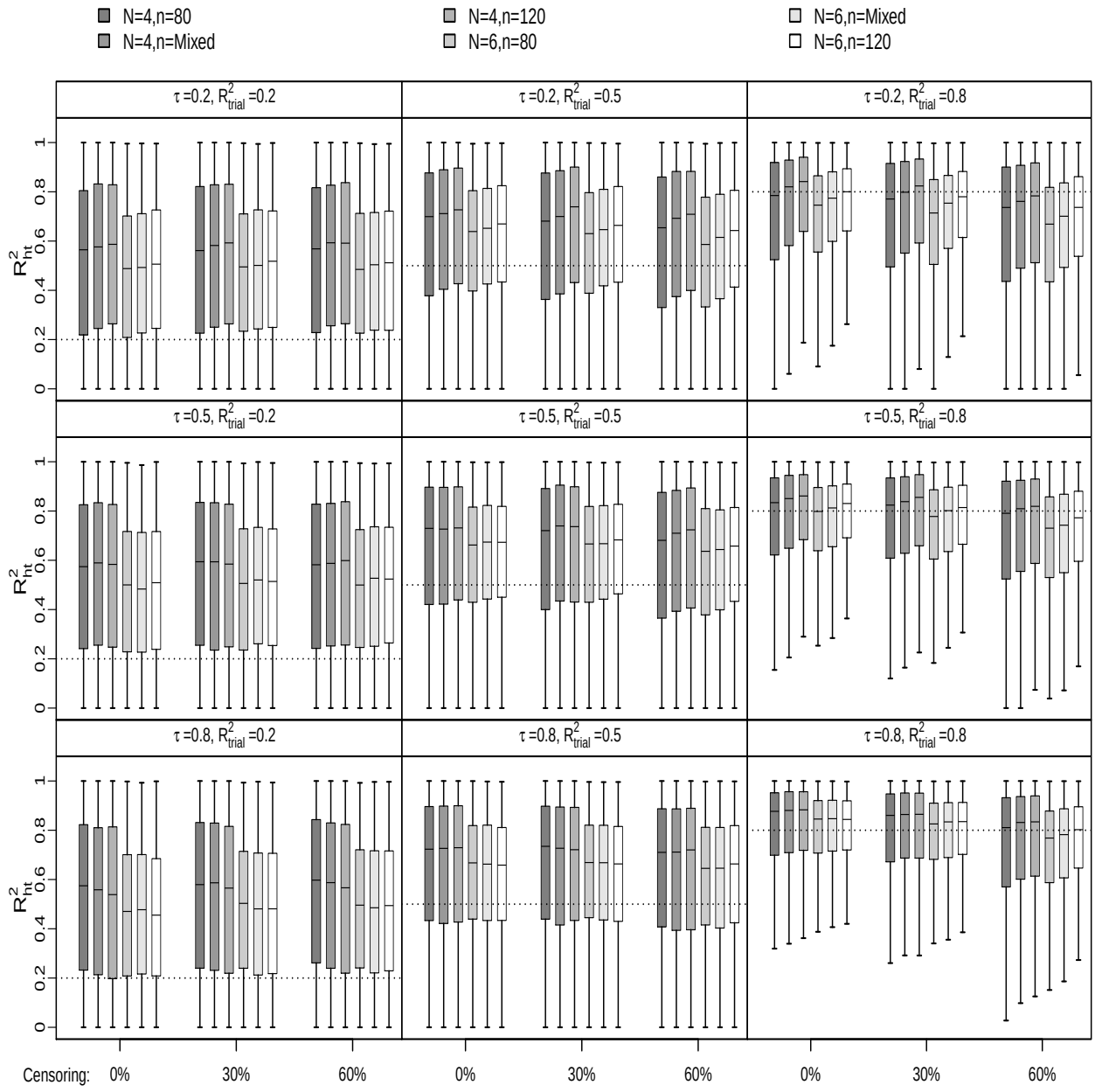

Figure B.9: Boxplots of estimates of R^2_{trial}: PFS, Gumbel Copula Data Generation, Information Theory Application (wider range of treatment effects on T)

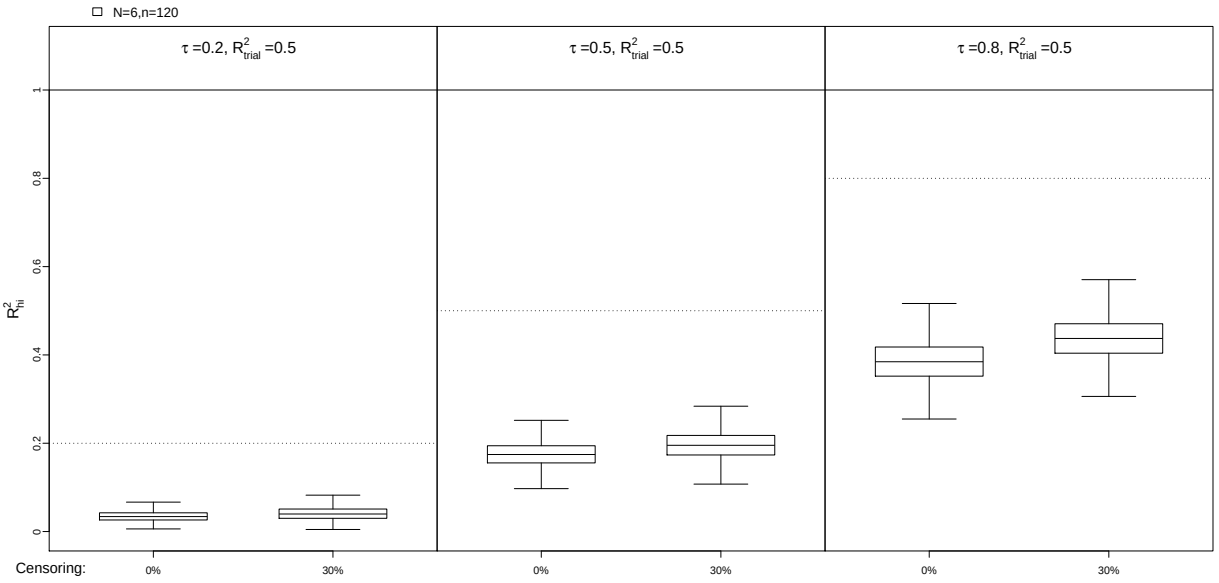

Figure B.10: Boxplots of estimates of $R^2_{h,i}$: TTP, Clayton Copula Data Generation, Information Theory Application (stronger treatment effects [HR 0.50 for PFS, HR 0.67 for OS])

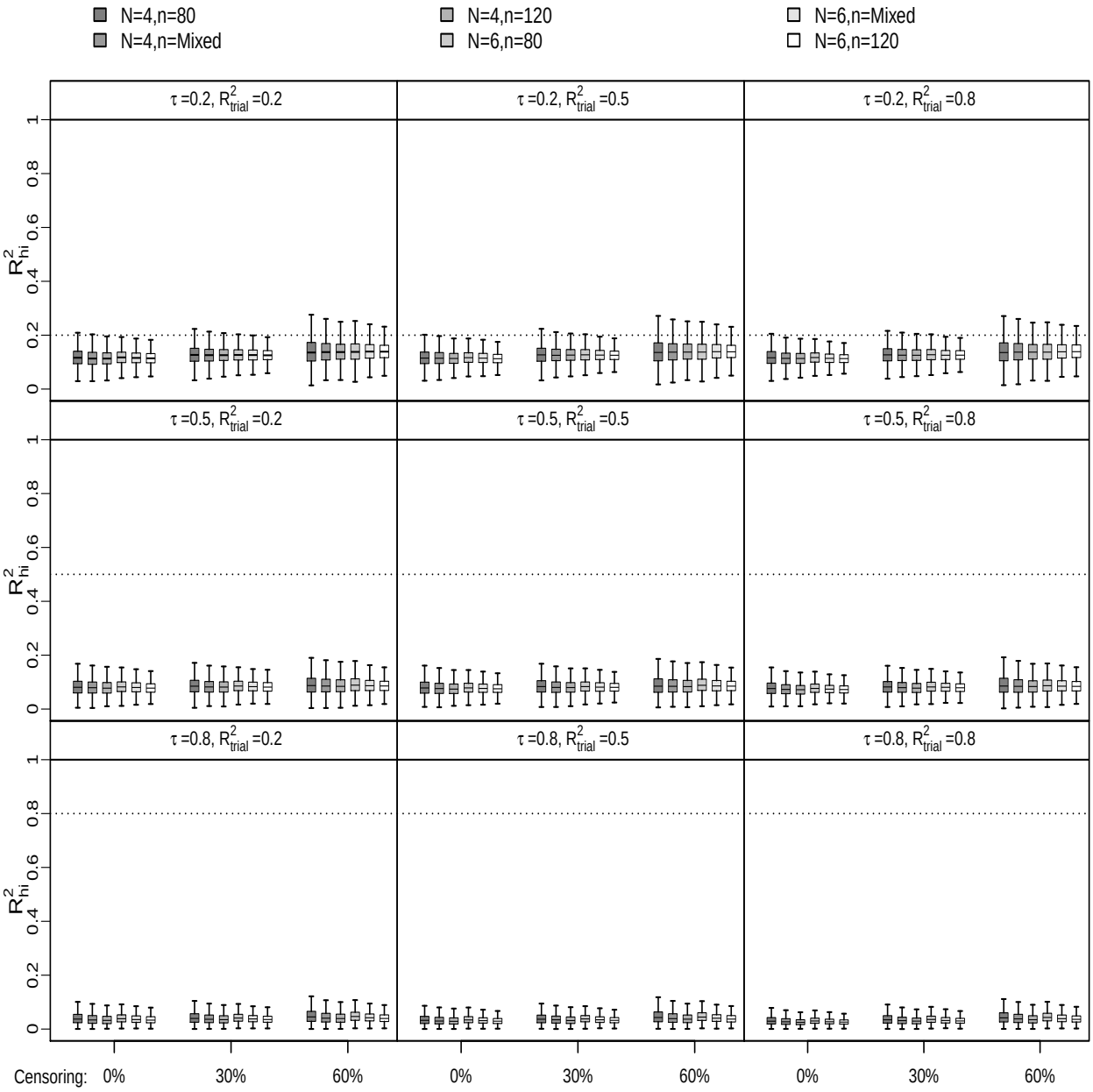

Figure B.11: Boxplots of estimates of $R^2_{h,i}$: TTP, Clayton Copula Data Generation, Information Theory Application (T-S)

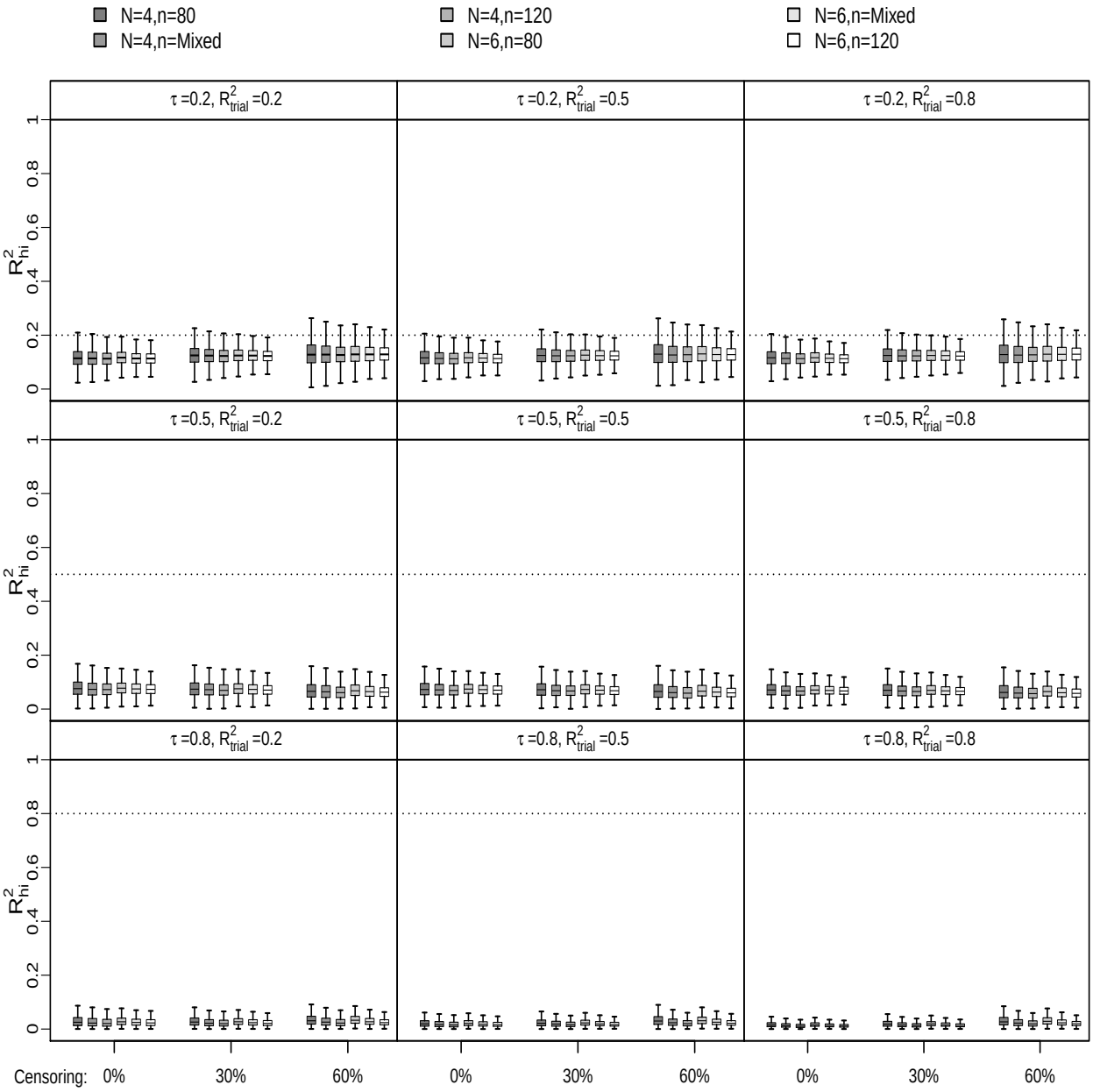

Figure B.12: Boxplots of estimates of $R^2_{h,i}$: TTP, Gumbel Copula Data Generation, Information Theory Application (T-S)

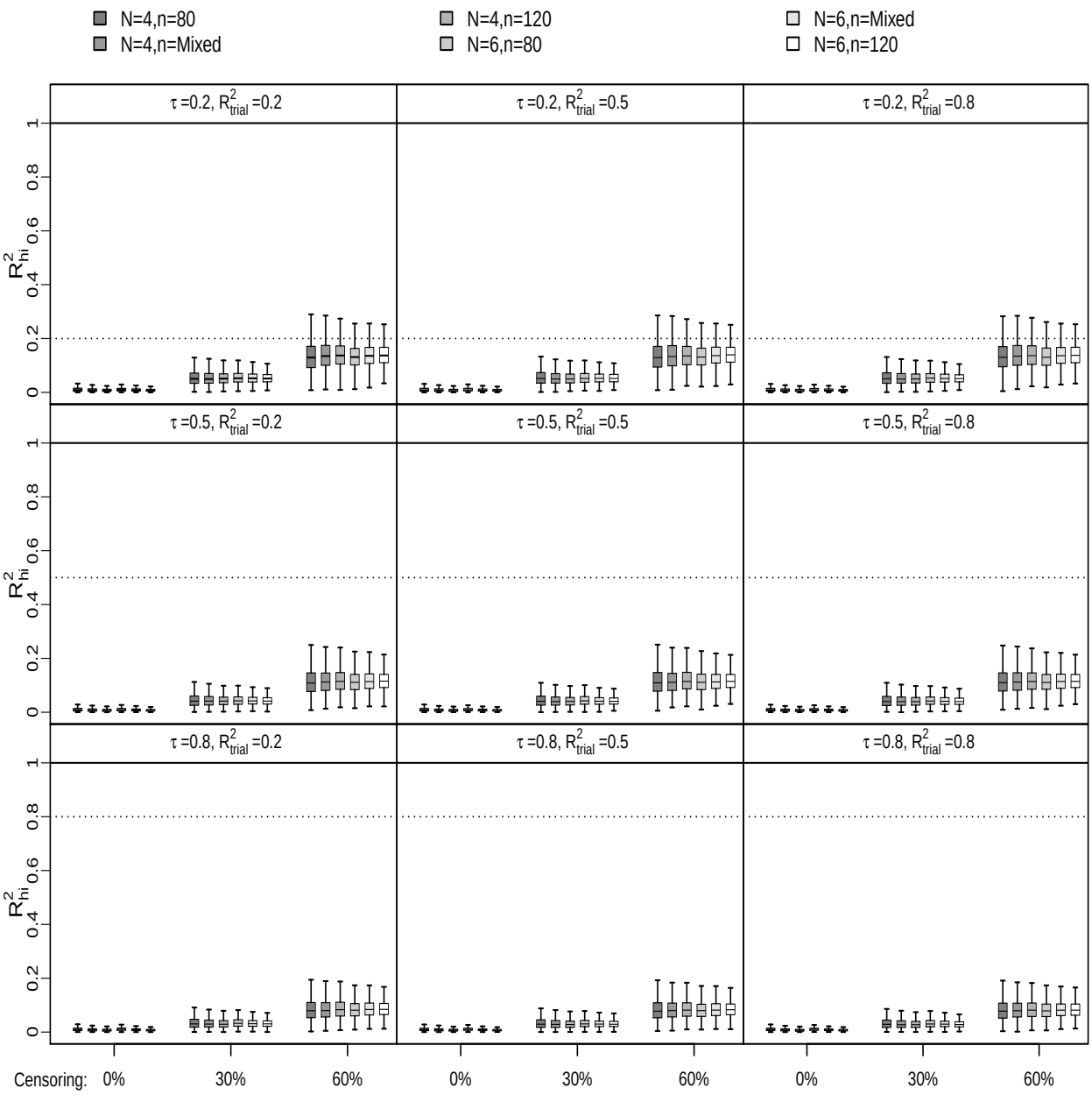

Figure B.13: Boxplots of estimates of $R^2_{h,i}$: PFS, Clayton Copula Data Generation, Information Theory Application (T-S)

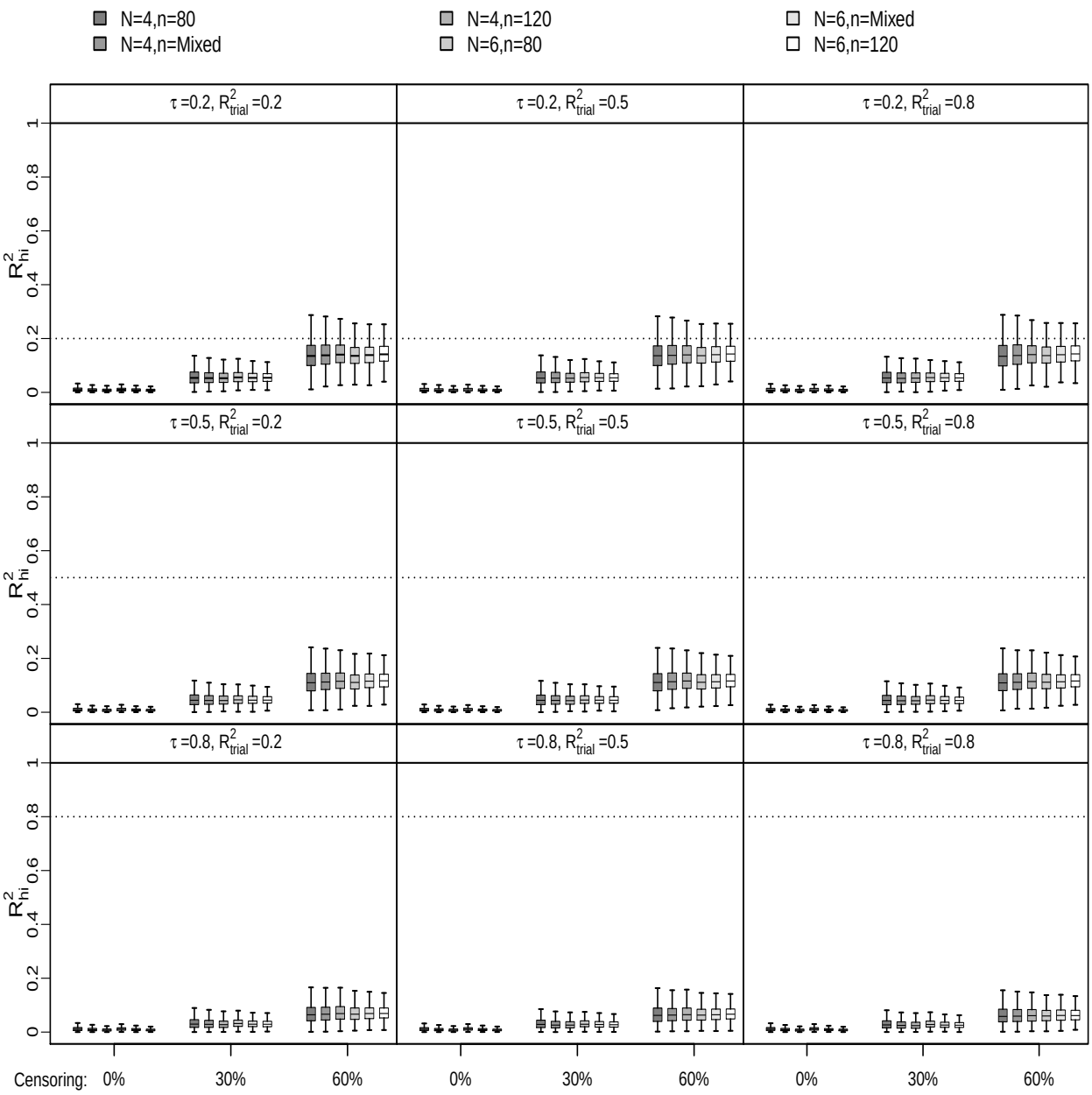

Figure B.14: Boxplots of estimates of $R^2_{h,i}$: PFS, Gumbel Copula Data Generation, Information Theory Application (T-S)

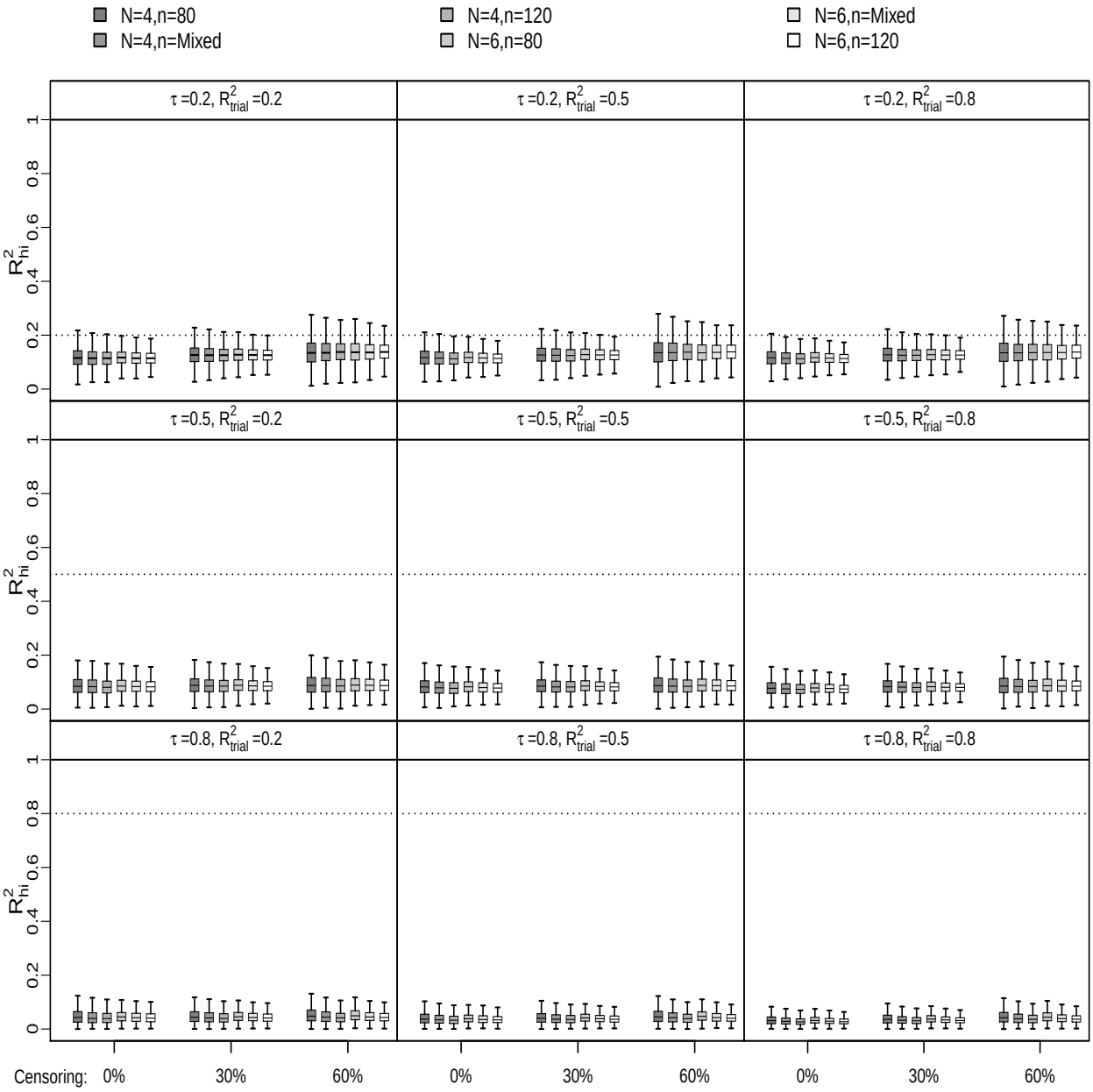

Figure B.15: Boxplots of estimates of $R^2_{h,i}$: TTP, Clayton Copula Data Generation, Information Theory Application (T-S) (wider treatment effects)

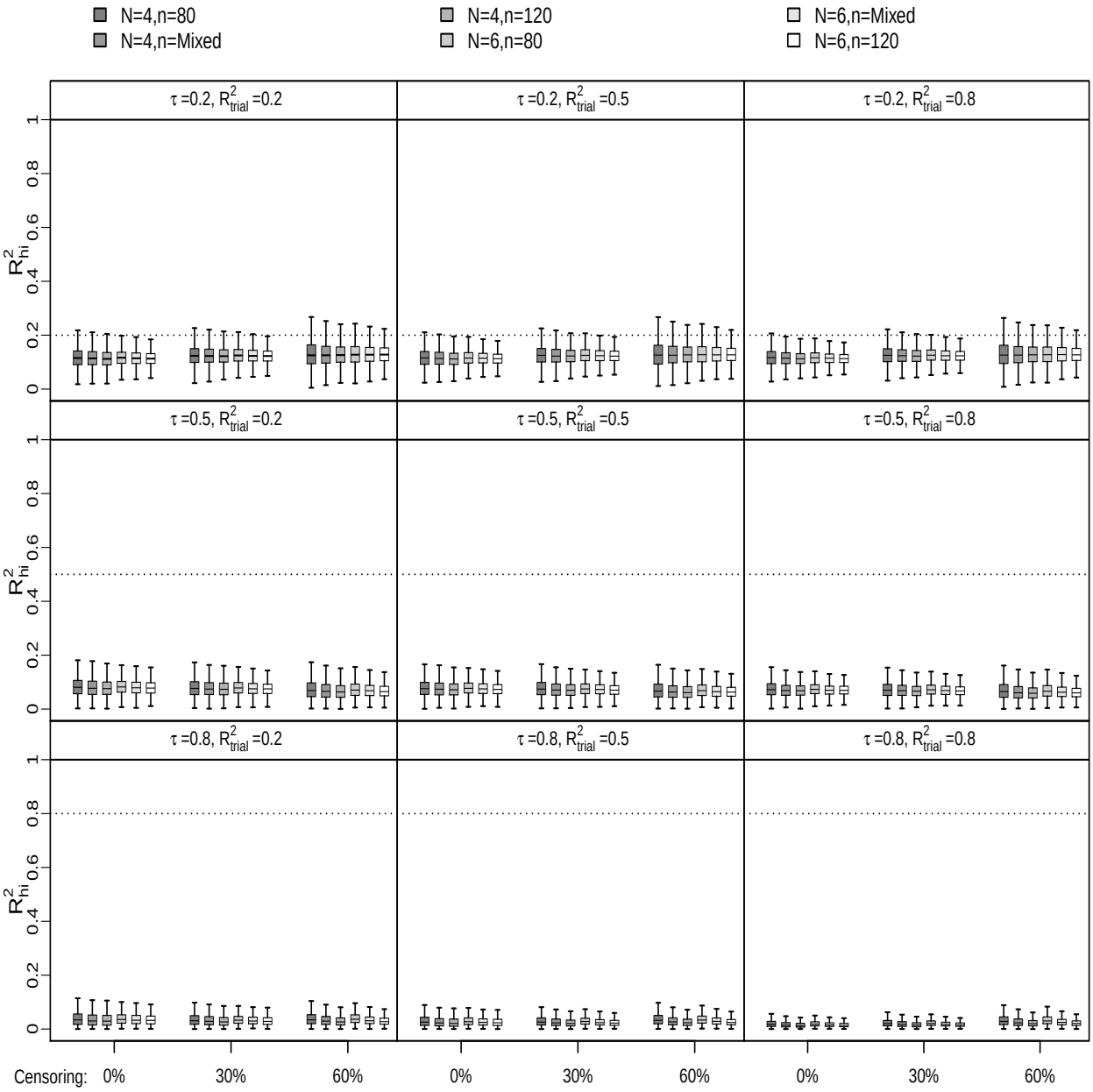

Figure B.16: Boxplots of estimates of $R^2_{h,i}$: TTP, Gumbel Copula Data Generation, Information Theory Application (T-S) (wider treatment effects)

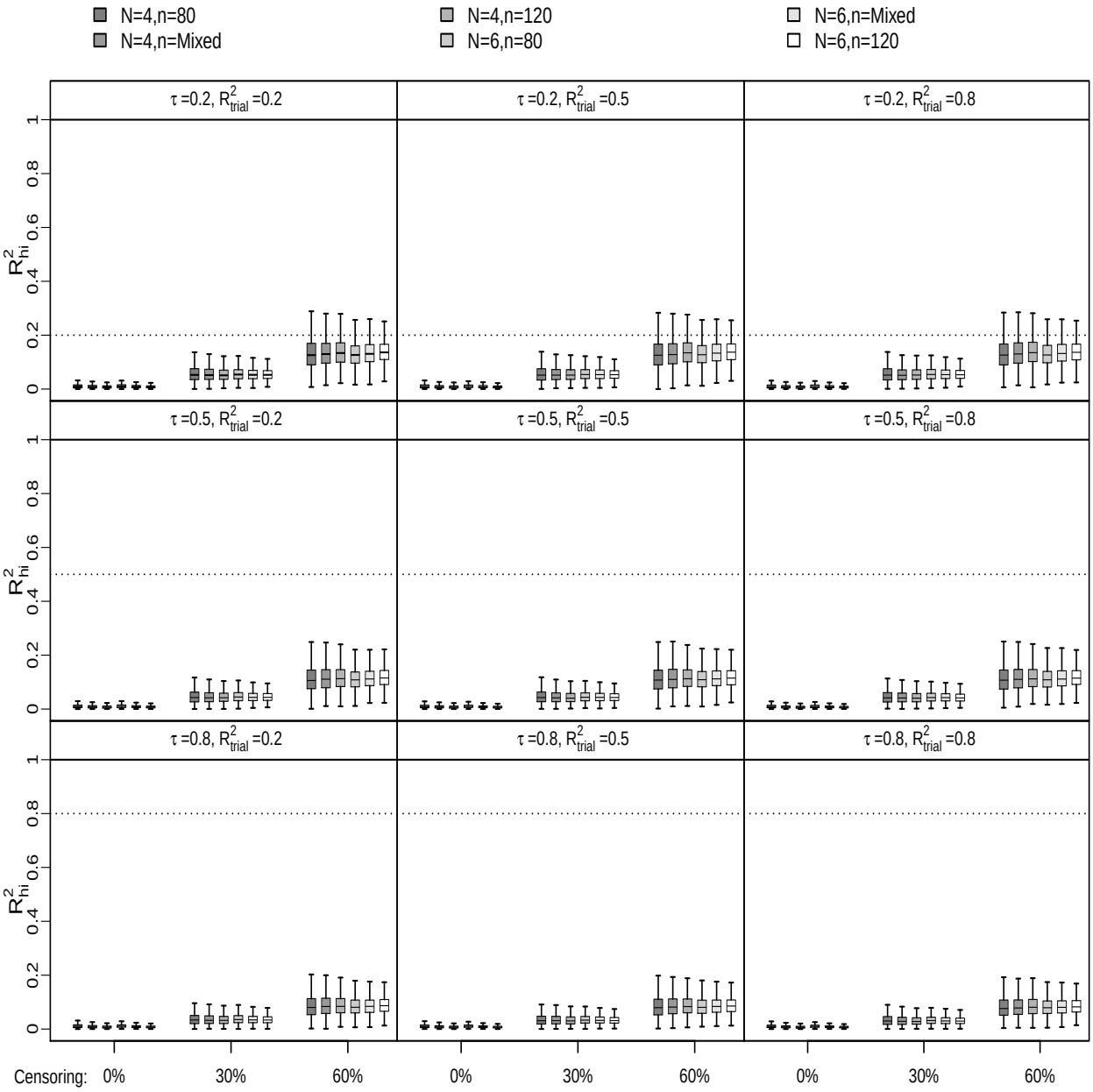

Figure B.17: Boxplots of estimates of $R^2_{h,i}$: PFS, Clayton Copula Data Generation, Information Theory Application (T-S) (wider treatment effects)

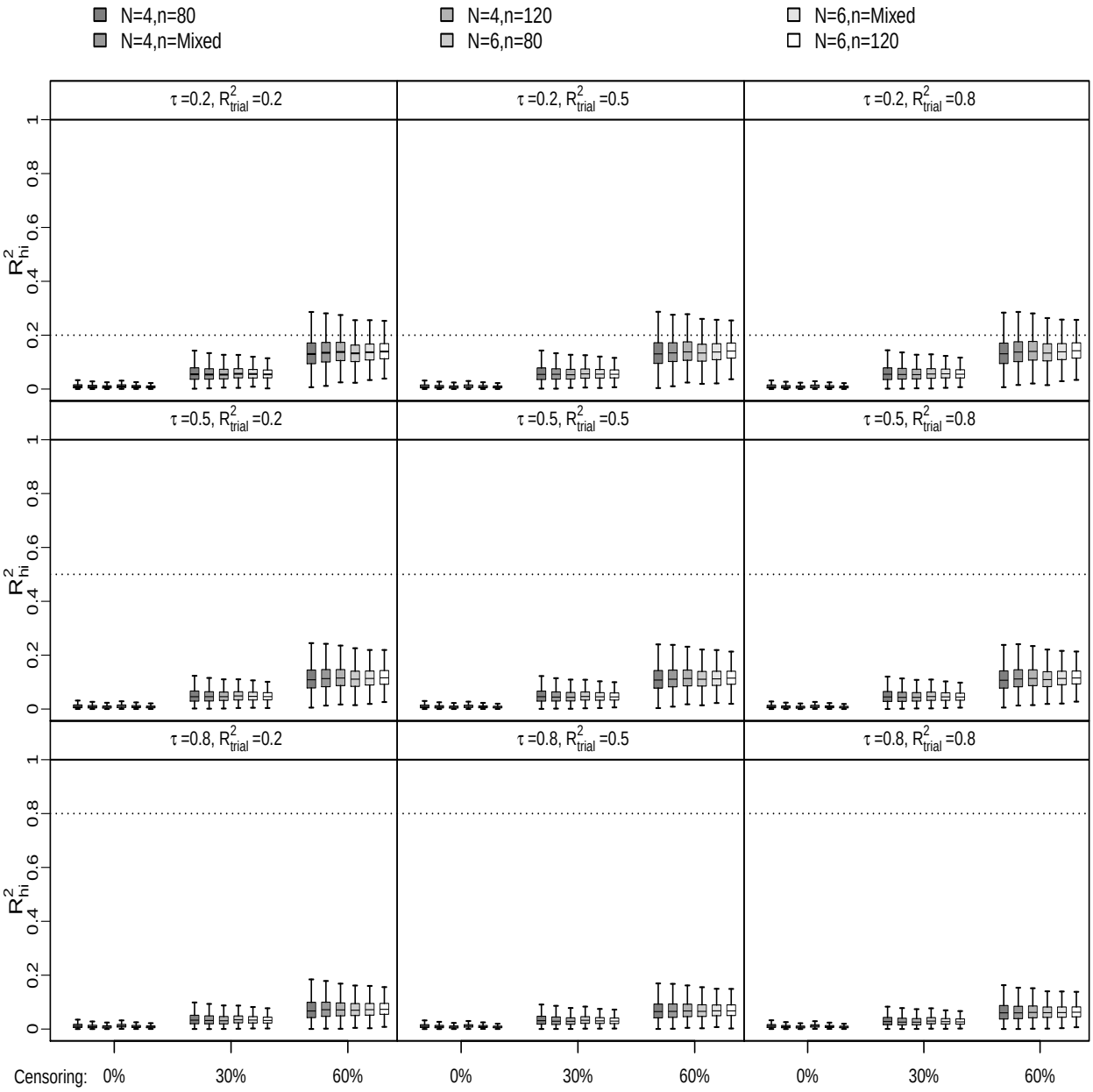

Figure B.18: Boxplots of estimates of $R^2_{h,i}$: PFS, Gumbel Copula Data Generation, Information Theory Application (T-S) (wider treatment effects)

Appendix C

Total Gain Method

Additional Results

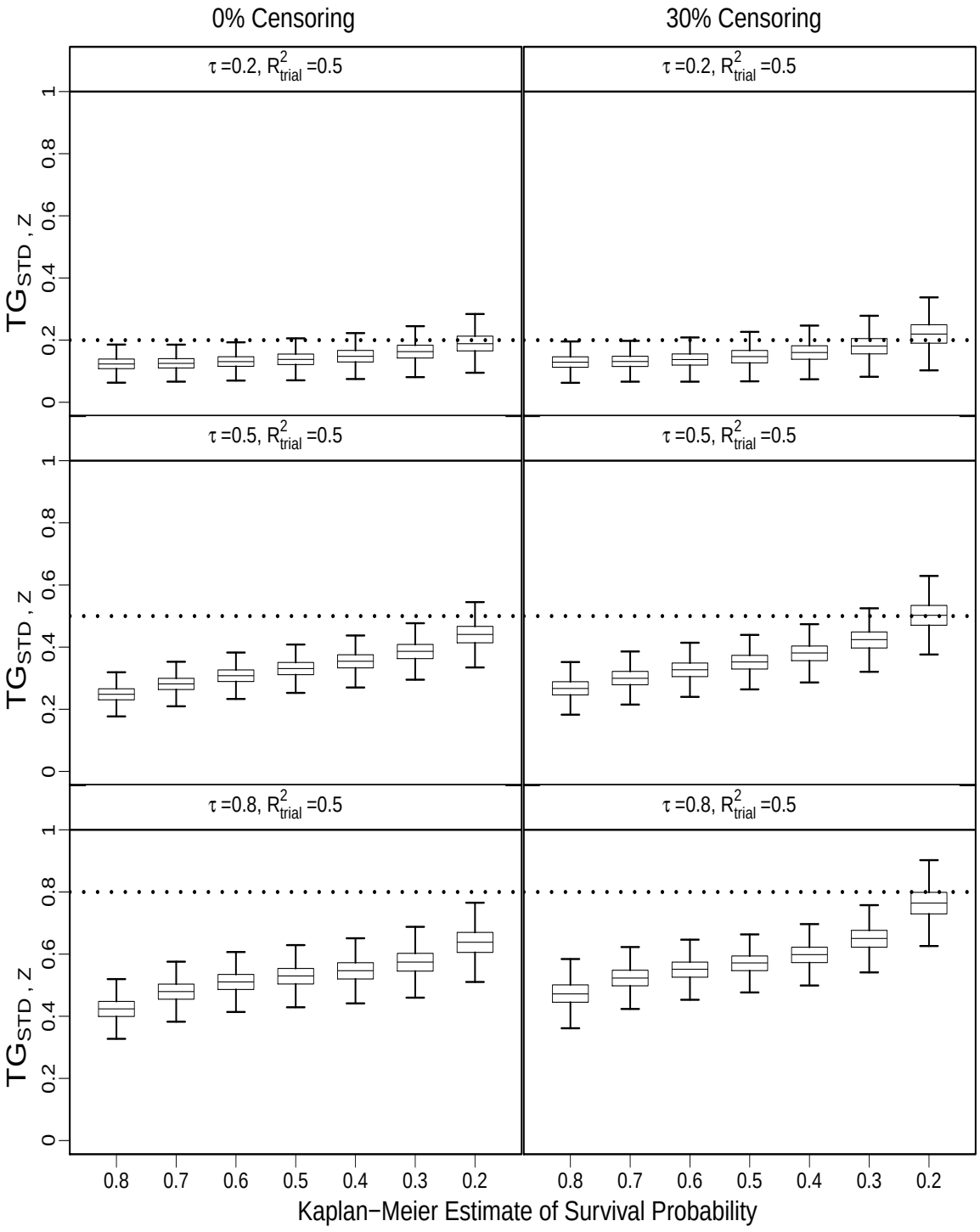

Figure C.1: Boxplots of estimates of $TG_{STD,Z}(t)$ at Percentiles of OS: TTP, Clayton Data Generation, Total Gain Application

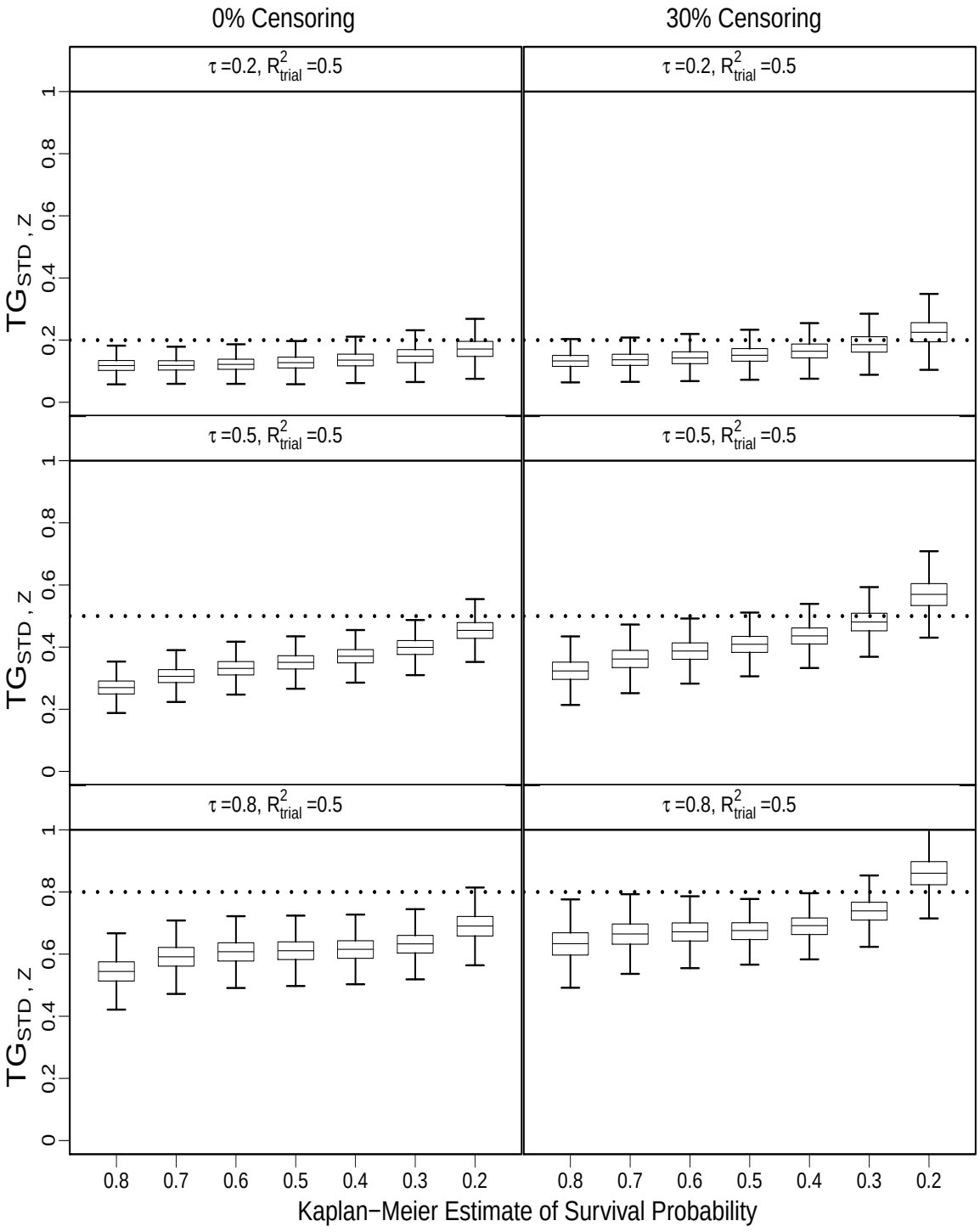

Figure C.2: Boxplots of estimates of $TG_{STD,Z}(t)$ at Percentiles of OS: TTP, Gumbel Data Generation, Total Gain Application

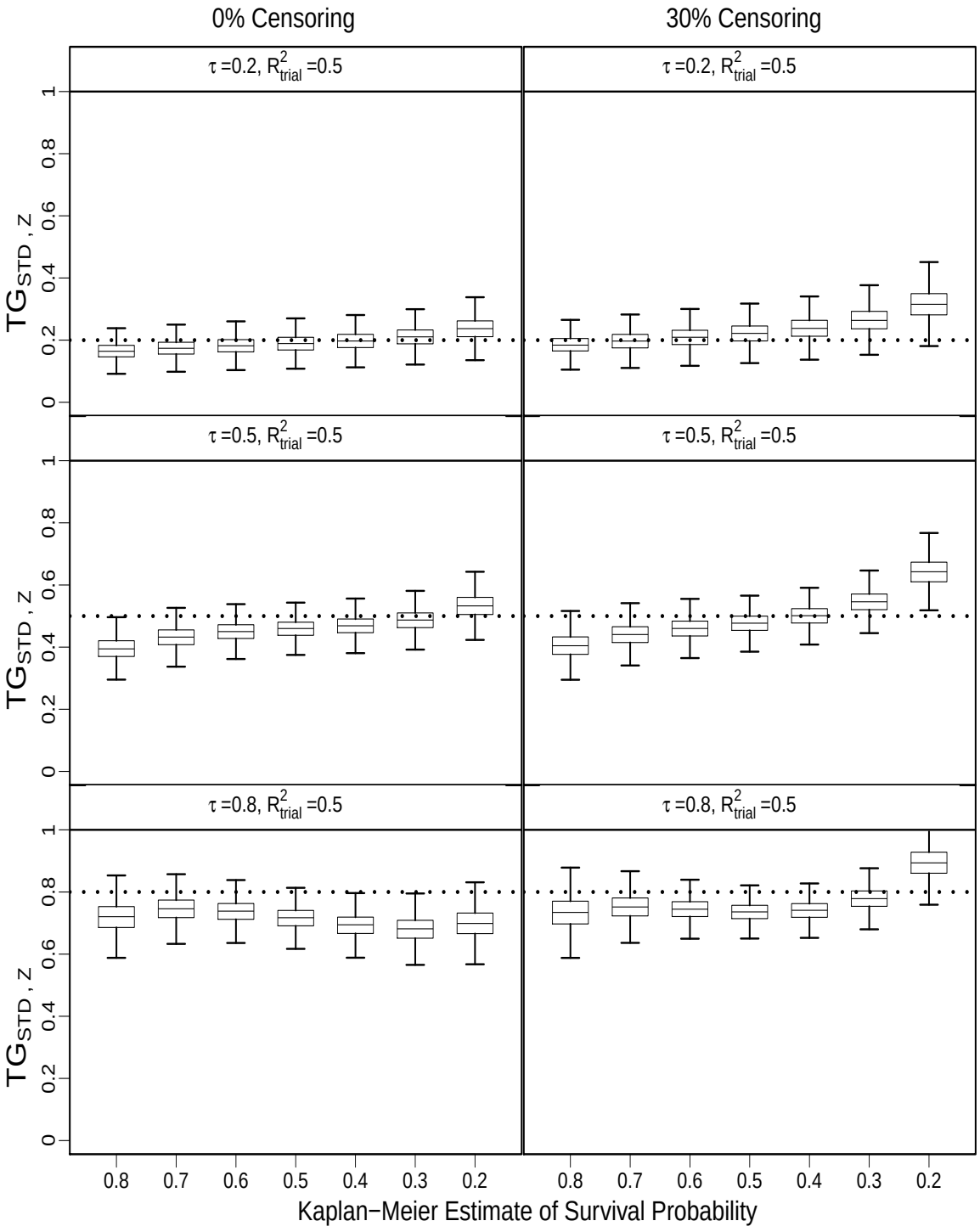

Figure C.3: Boxplots of estimates of $TG_{STD,Z}(t)$ at Percentiles of OS: PFS, Clayton Data Generation, Total Gain Application

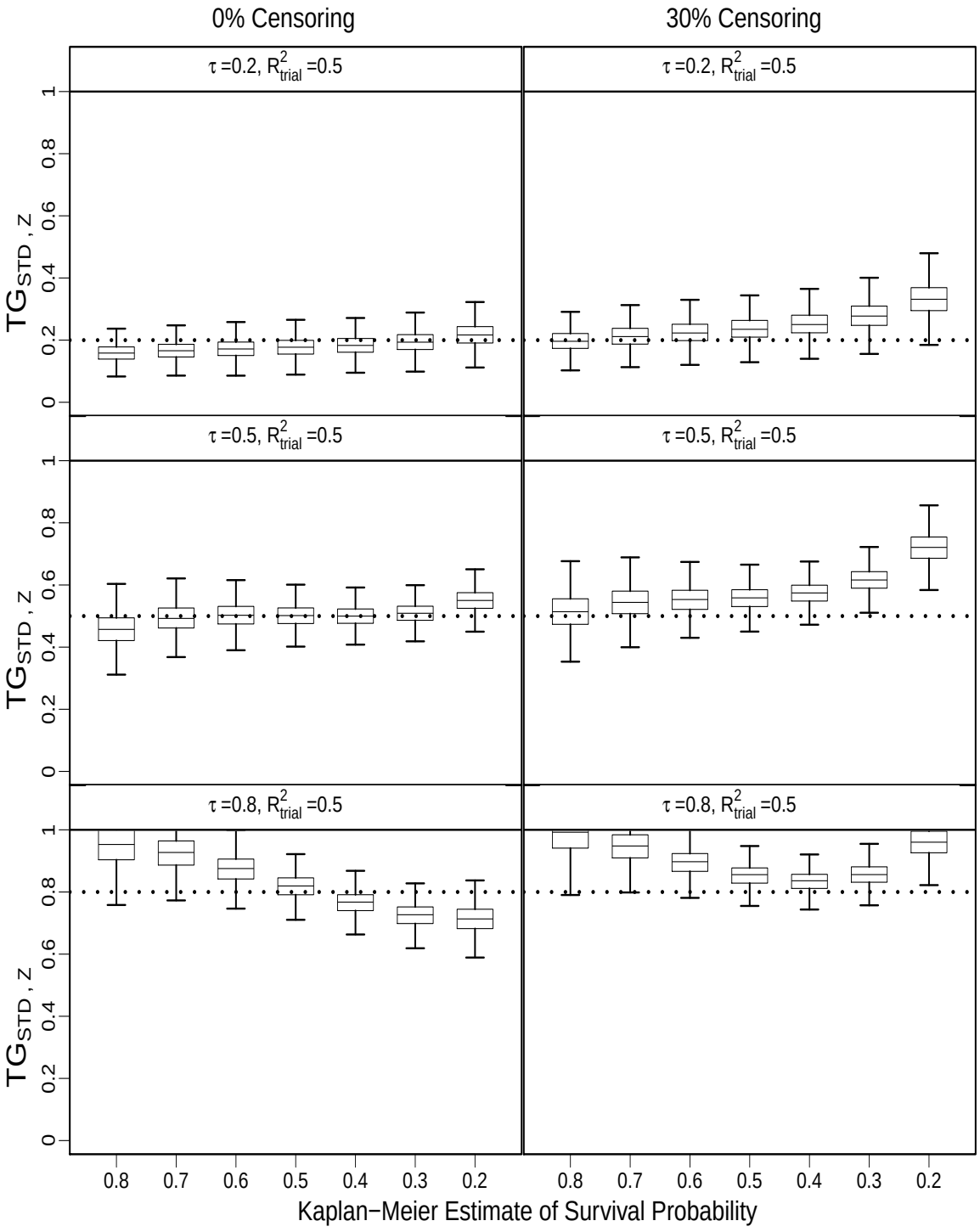

Figure C.4: Boxplots of estimates of $TG_{STD,Z}(t)$ at Percentiles of OS: PFS, Gumbel Data Generation, Total Gain Application

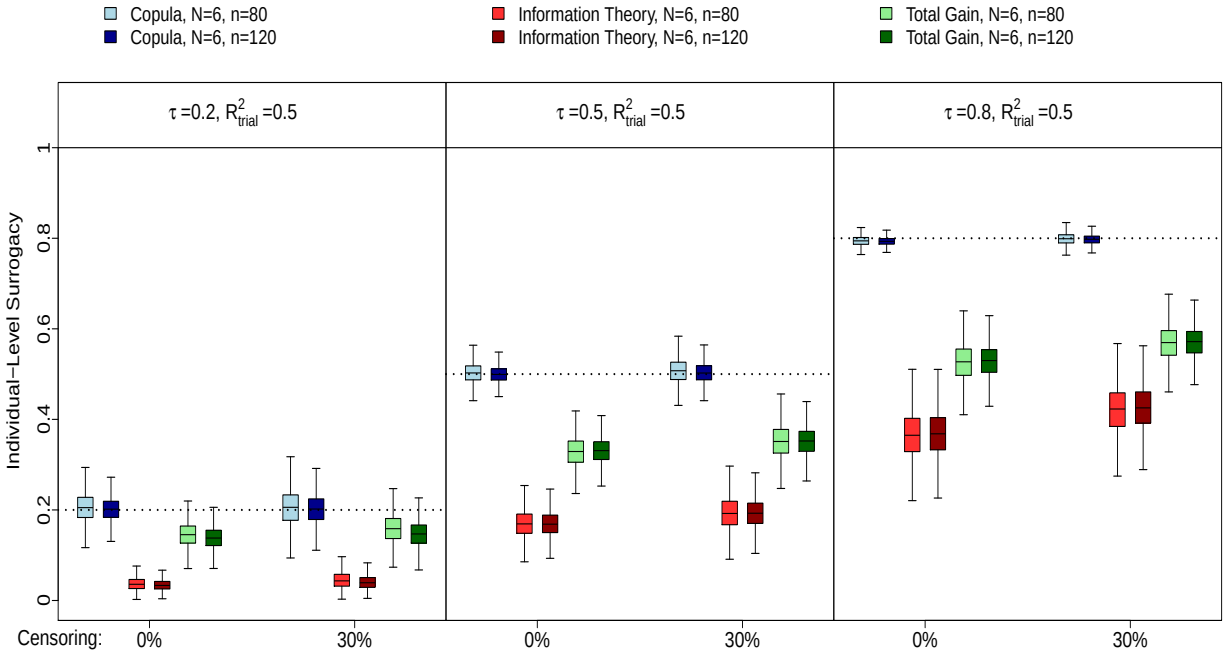

Figure C.5: Boxplots of all surrogacy methods: TTP, Clayton Copula Data Generation

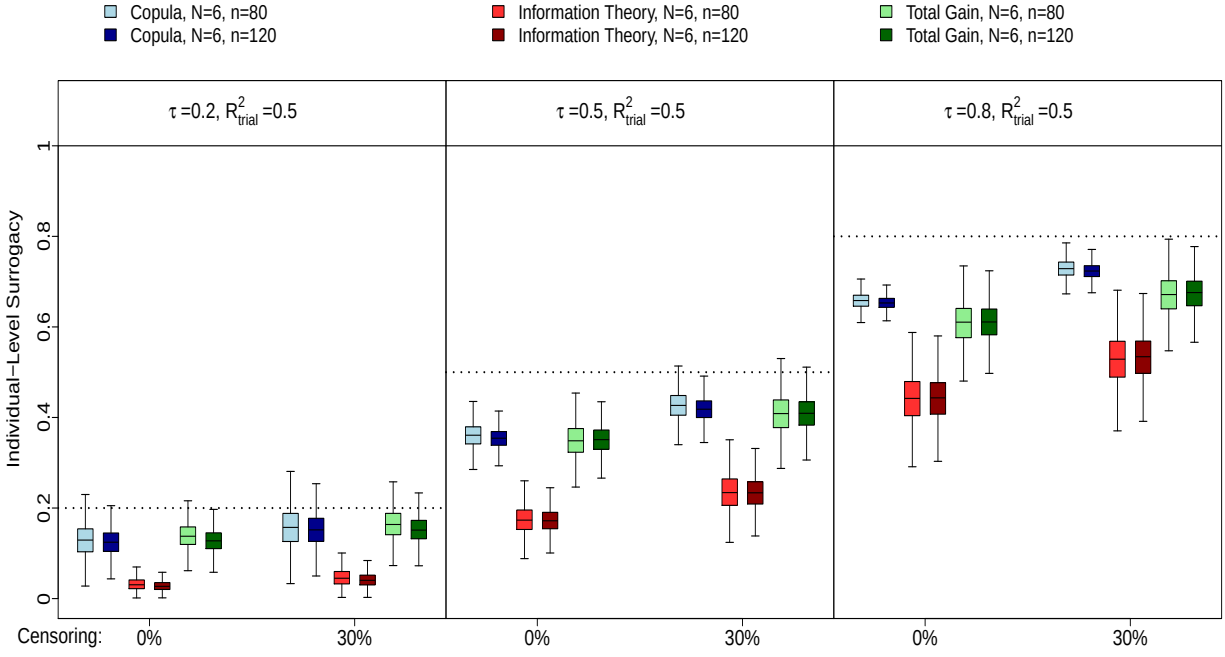

Figure C.6: Boxplots of all surrogacy methods: TTP, Gumbel Copula Data Generation

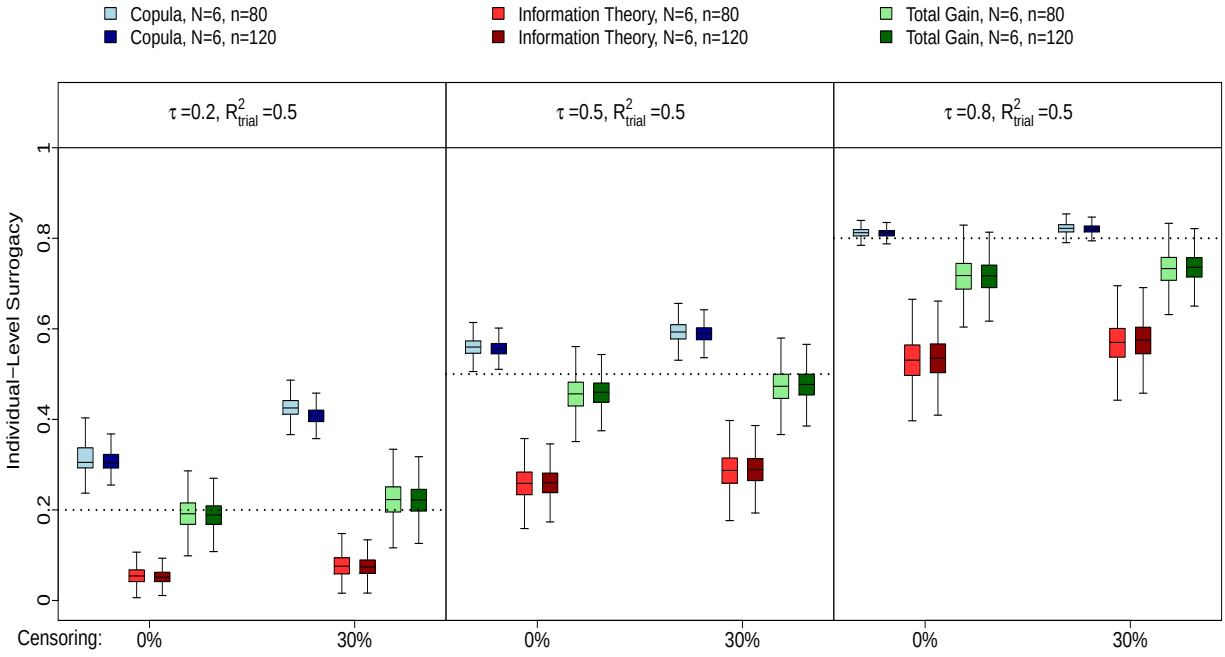

Figure C.7: Boxplots of all surrogacy methods: PFS, Clayton Copula Data Generation

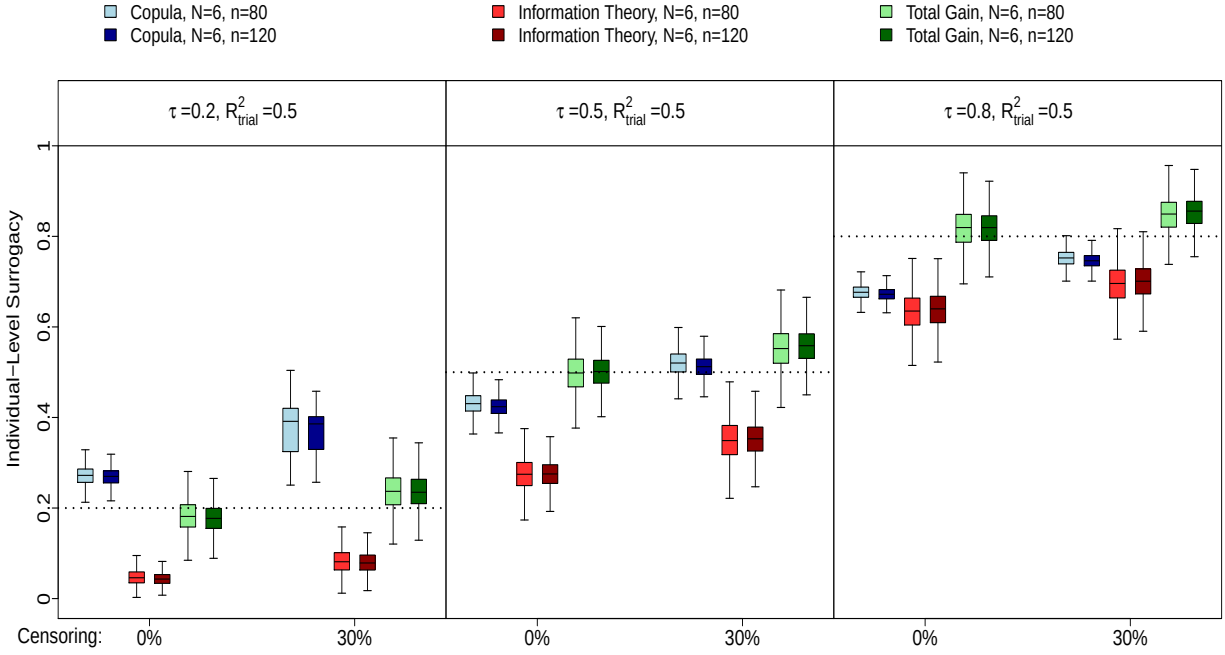

Figure C.8: Boxplots of all surrogacy methods: PFS, Gumbel Copula Data Generation

Appendix D

Publication

Received: 15 February 2016 | Revised: 30 January 2017 | Accepted: 14 April 2017

DOI: 10.1002/pst.1812

MAIN PAPER

WILEY

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[1,2] | Susan Todd[2]

[1] Roche Products Ltd, Welwyn Garden City, UK

[2] Department of Mathematics and Statistics, University of Reading, Reading, UK

Correspondence
Natalie Dimier, Roche Products Ltd, Hexagon Place, 6 Falcon Way, Shire Park, Welwyn Garden City AL7 1TW, UK.
Email: natalie.dimier@roche.com

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
meta-analysis, oncology, progression-free survival, surrogate endpoint, time to progression

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 health care. 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 (OS), being objective, reliable, and easy to measure. However, demonstrating a clinical benefit in survival is becoming increasingly complex because of 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 OS 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 the so-called surrogate endpoints.[1]

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). As a result of 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 regarding methodology for evaluating surrogate endpoints. These include single-trial hypothesis testing methods,[2] approximation methods,[3–7] and 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 Buyse et al[20] and Laporte et al[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 on the basis of modelling the joint survival distribution of the 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 OS in oncology studies. In reality, 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. Progression-free survival 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 semicompeting 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 (TTP and PFS) as surrogates for OS 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

To thoroughly assess a potential surrogate endpoint, Burzykowski et al[25] recommend to explore 2 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, 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 that relates the underlying dependence between them. Surrogacy is evaluated through a two-stage procedure, where stage one fits the copula to the data 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 $j = 1, \ldots, n$ patients 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

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

A value of R^2_{trial} close to 1 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 0 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. 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 because of nonconvergence 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-level 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 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

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 OS, 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, 0.60-0.91; median OS of 13.8 versus 11.1 months in the experimental and control arms) was sufficient to cross the prespecified 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 versus 5.5 months and hazard ratio 0.71 [95% confidence interval, 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 substudy within the trial. Further discussion of this approach can be found in Renfro et al.[24] Countries containing 7 or fewer patients were grouped by geographical region to allow for parameter estimation; 2 countries were removed from analysis because of 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.

There are a number of aims of our study; the first is to determine how well the method performs when using a surrogate end point that combines multiple events of interest, including the event of interest for the true endpoint. In the original simulation

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

Abbreviations: PFS, progression-free survival; TTP, time to progression.

study performed by Burzykowski et al,[12] the simulated data are constructed according to a TTP 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 TTP and 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 patients, 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 to 6 clinical trials containing 80 to 120 patients 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-level and trial-level association can be accurately controlled. 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 to assess the impact of model misspecification. These two copula functions assume different underlying dependence structures of the 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 because of 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 5000 per scenario. Simulations were run on a Windows 7 64-bit machine with 4 GB 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 end point, 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,[23] the marginal survival functions are chosen to follow an exponential survival distribution. As described by

Burzykowski,[23] trial-specific random effects are used to control the trial-level association. 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 Burzykowski[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), to reflect the potential influence of postprogression 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 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.

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 end points. 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 to achieve true individual-level association of 0.2, 0.5, and 0.8, respectively. 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 Burzykowski.[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. 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 (ie, earlier event times), whereas the Clayton exhibits strong upper-tail dependence (ie, later event times). For the two endpoints, S and T, the form of the Gumbel model is

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

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.

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

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

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

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

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

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 $\widetilde{S}_{ij}^{0}$ and $\widetilde{T}_{ij}^{0}$ 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 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 patients 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 patients to be censored for the surrogate earlier than the time of censoring of OS, representing scenarios where patients 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 nonconvergence rate of 1.12%, most of which occurred for low levels of true individual-level association. However, when PFS was used as the surrogate, nonconvergence was significantly worse, reaching as high as 61.3% for low individual association. In both cases, the nonconvergence 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 nonconvergence, 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 (4 trials of 80 patients). Results of larger sample sizes can be found in the Supporting Information. Additionally, since there was little variation in results with varying true underlying trial-level association, the results presented here represent only scenarios with $R^2_{\text{trial}} = 0.5$. Results for varying values of R^2_{trial} can also be found in the Supporting Information. Horizontal dashed lines at $y = 0.2, 0.5, 0.8$ represent the true individual-level surrogacy being estimated by each set of 3 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 underestimation under the presence of little to no censoring. Overall, the maximum average relative bias is −38.1%, demonstrating that the method most often underestimates 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 that 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 OS. 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

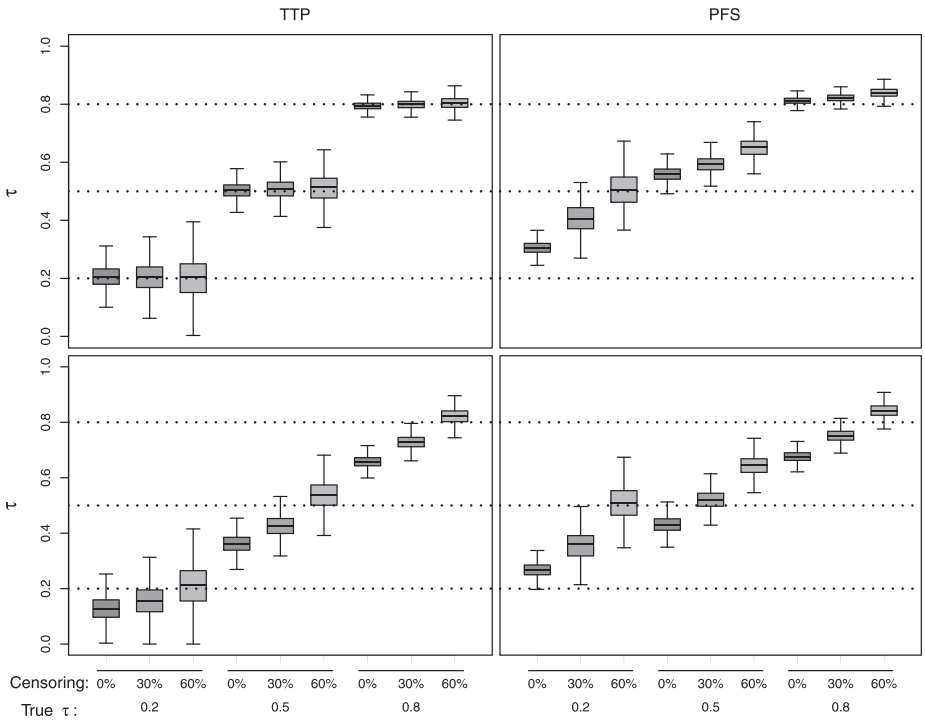

FIGURE 1 Estimated values of τ

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 (4 trials with 80 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 underestimated. 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-level and trial-level association, with increasing R^2_{trial} estimates with increased true individual association. To verify results of previous simulation studies conducted 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

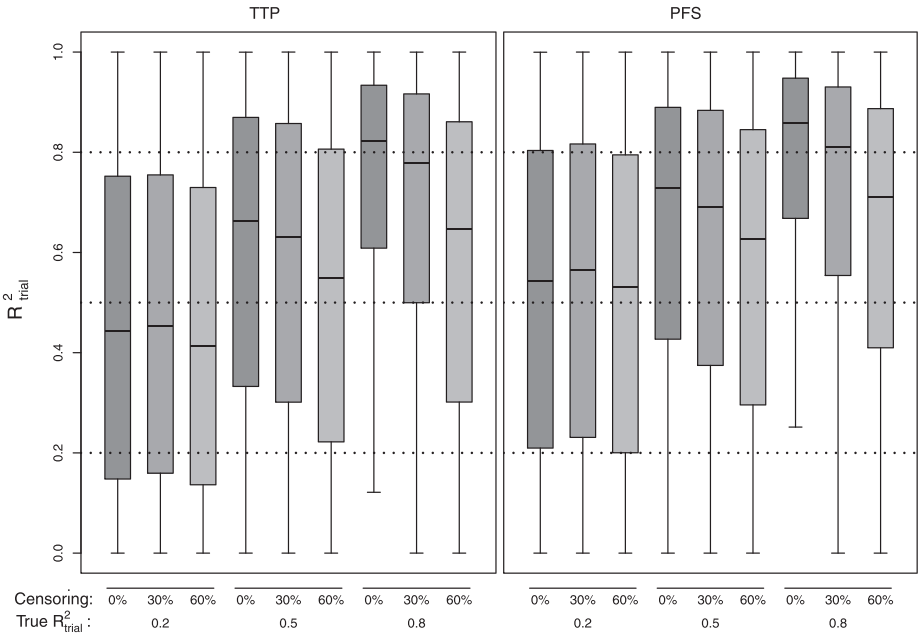

FIGURE 2 Estimated values of R^2_{trial}

surrogate endpoints. A large range of scenarios were considered, including varied sample sizes, varying strength of individual-level 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 TTP endpoint, with the level of variability reflecting the small sample sizes used in this study. The change in underlying data structure led to slight underestimation, 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 patients 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 (eg, 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 overestimation 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 on the basis of 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 overestimation 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 OS 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 that need to be explored further. The work of Renfro et al[30] suggests that alternative modelling using a 2-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 multistate model.[35] As discussed previously, a semicompeting 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 to 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 6 trials, a measure of R^2_{trial} based on trials as units does not provide reliable conclusions. Additionally, even when there are only 4 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 TTP is performed using this method. As noted by Burzykowski[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.

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] J. R. Johnson, Y. M. Ning, A. Farrell, R. Justice, P. Keegan, R. Pazdur, *J. Natl. Cancer Inst.* **2011**, *103*, 1–9.

[2] R. Prentice, *Stat. Med.* **1989**, *8*, 431–440.

[3] L. S. Freedman, B. I. Graubard, A. Schatzkin, *Stat. Med.* **1992**, *11*, 167–178.

[4] D. Y. Lin, T. R. Fleming, V. DeGruttola, *Stat. Med.* **1997**, *16*, 1515–1527.

[5] M. K. Cowles, *Stat. Med.* **2002**, *21*, 811–834.

[6] Y. Wang, J. M. G. Taylor, *Biometrics* **2002**, *58*, 803–812.

[7] Z. Li, M. P. Meredith, M. S. Hoseyni, *Stat. Med.* **2001**, *20*, 3175–3188.

[8] M. J. Daniels, M. D. Hughes, *Stat. Med.* **1997**, *16*, 1965–1982.

[9] M. Buyse, G. Molenberghs, *Biometrics* **1998**, *54*, 1014–1029.

[10] M. Buyse, G. Molenberghs, T. Burzykowski, D. Renard, H. Geys, *Biostatistics* **2000**, *1*, 49–67.

[11] M. H. Gail, R. Pfeiffer, H. C. van Houwelingen, R. J. Carroll, *Biostatistics* **2000**, *1*, 231–246.

[12] T. Burzykowski, G. Molenberghs, M. Buyse, H. Geys, D. Renard, *Appl. Stat.* **2001**, *50*, 405–422.

[13] F. Tibaldi, F. T. Barbosa, G. Molenberghs, *Stat. Med.* **2004**, *23*, 2173–2186.

[14] A. Alonso, G. Molenberghs, T. Burzykowski, D. Renard, H. Geys, Z. Shkedy, F. Tibaldi, J. Cortinas Abrahantes, M. Buyse, *Biometrics* **2004**, *60*, 724–728.

[15] T. Burzykowski, M. Buyse, *Pharm. Stat.* **2006**, *5*, 173–186.

[16] A. Alonso, G. Molenberghs, *Biometrics* **2007**, *63*, 180–186.

[17] A. Pryseley, A. Tilahun, A. Alonso, G. Molenberghs, *Lifetime Data Anal.* **2011**, *17*, 195–214.

[18] C. J. Weir, R. J. Walley, *Stat. Med.* **2006**, *25*, 183–203.

[19] H. Ensor, R. Lee, C. Sudlow, C. J. Weir, *J. Biopharm. Stat.* **2015**, *26*(5), 859–879.

[20] M. Buyse, S. Michiels, P. Squifflet, K. J. Lucchesi, K. Hellstrand, M. L. Brune, S. Castaigne, J. M. Rowe, *Haematologica* **2011**, *96*, 1106–1112.

[21] S. Laporte, P. Squifflet, N. Baroux, F. Fossella, V. Georgoulias, J. L. Pujol, J. Y. Douillard, S. Kudoh, J. P. Pignon, E. Quinaux, M. Buyse, *BMJ Open* **2013**, *3*, e001802, https://doi.org/10.1136/bmjopen-2012-001802.

[22] D. Ghosh, J. M. G. Taylor, D. J. Sargent, *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] L. A. Renfro, Q. Shi, Y. Xue, J. Li, H. Shang, D. J. Sargent, *Comput. Stat. Data Anal.* **2014**, *78*, 1–20.

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

[26] M. Kendall, *Biometrika* **1938**, *30*(1–2), 81–93.

[27] H. C. van Houwelingen, L. R. Arends, T. Stijnen, *Stat. Med.* **2002**, *21*, 589–624.

[28] L. Renfro, Q. Shi, D. Sargent, B. Carlin, *Stat. Med.* **2012**, *31*, 743–761.

[29] https://ibiostat.be/software/surrogate (accessed 6th February 2016).

[30] L. A. Renfro, H. Shang, D. J. Sargent, *J. Biopharm. Stat.* **2015**, *25*, 857–877.

[31] Y. J. Bang, E. Van Cutsem, A. Feyereislova, H. C. Chung, L. Shen, A. Sawaki, F. Lordick, A. Ohtsu, Y. Omuro, T. Satoh, G. Aprile, E. Kulikov, J. Hill, M. Lehle, J. Rschoff, Y. K. Kang, *Lancet* **2010**, *376*, 687–697.

[32] Copyright, SAS Institute Inc. Cary, NC, USA.

[33] R. G. Nelsen, *An Introduction to Copulas*, Springer Verlag **1999**.

[34] P. K. Trivedi, D. M. Zimmer, *Found. Trends Econom.* **2005**, *1*, 1–111.

[35] D. DeJardin, E. Lesaffre, G. Verbeke, *Stat. Med.* **2010**, *29*, 1724–1734.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: 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. https://doi.org/10.1002/pst.1812