THE UNIVERSITY
of EDINBURGH

Statistical evaluation of surrogate outcomes: methodological extensions to ordinal outcomes with applications in acute stroke

Hannah Margaret Ensor

Doctorate - The University of Edinburgh - 2016

Contents

Table of Figures

Table of tables

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or processional qualification.

The first stage of the systematic review search process, as clearly outlined in Chapter 2, was carried out by Robert Lee (this was conducted previous to the start of my studentship). A review of all papers after initial screening (at all stages) was carried out by myself, Robert Lee or my supervisor Christopher Weir. The systematic review chapter is pre-published online by the Journal of Biopharmaceutical Statistics (Ensor et al., 2015).

The clinical trial CLOT3 data for the case studies presented in Chapter 9 was provided by Prof Martin Dennis and Cat Graham. They also and gave guidance on clinical questions of interest.

Acknowledgements

I would like to thank my supervisors Christopher Weir and Cathie Sudlow. Cathie thank you for your invaluable input and our many thought provoking meetings. Chris, I am indebted to you for your excellent advice, encouragement and support. You always gave me feedback long before I could hope to expect it and easily guided me through the tougher aspects of the PhD journey. I do not see how anyone could have a better supervisor.

I would also like to thank Robert Lee for our many interesting and helpful conversations.

Thanks too to Cat Graham and Martin Dennis and everyone involved in the CLOTS3 trial.

Finally, thanks to my parents for the 'you are not trying to split the atom' advice and my partner Euan for the 'best eleven pages of my thesis'; and in both cases for much more besides.

Abstract

Background

Surrogate outcomes are measures of treatment effect that can be used to predict treatment effect on the true outcome of interest. Surrogates are valued as they can be used in place of true outcomes to reduce the length, size, or intrusiveness of a clinical trial. However, validation of surrogacy is a conceptually complicated area and much theoretical and practical statistical development has been conducted in recent years.

Methods

A systematic review was conducted to identify which surrogate evaluation approach was best suited to be extended to ordinal outcomes. I extended a foremost approach to the case where the surrogate, the true clinical outcome, or both are ordinal outcomes. This extension investigated surrogacy at both the trial and individual levels; trial level surrogacy was based on a two stage method. The extension was developed through large simulation studies and used to investigate whether deep venous thromboembolism (DVT) was a surrogate for the ongoing measure of death and disability the Oxford Handicap Scale (OHS), using data from the stroke trial CLOTS3. CLOTS3 was a large multi-centre randomised clinical trial which investigated whether intermittent pneumatic compression (IPC) applied to the legs reduced the occurrence of deep venous thromboembolism (DVT) in stroke clinical trial patients.

Results

The systematic review identified the information theory approach as the most intuitively and practically worthwhile approach to surrogacy evaluation. I extended this approach to: a binary surrogate and ordinal true outcome (the binary-ordinal setting); the ordinal-binary and the ordinal-ordinal settings. The simulation studies showed that the approach worked well in most scenarios tested. However, trial level surrogacy was impacted by loss of efficiency due to the use of the two stage method. Bias imposed at the trial level by separation of discrete outcomes was effectively dealt with using a penalised likelihood method. The information theory approach for ordinal outcomes identified no surrogate that would predict treatment effect of IPC on the true outcome OHS measured at six months in the stroke trial CLOTS3.

Lay Summary

Imagine you have a serious illness with no known cure or treatment. Now suppose there is a drug that may help but no one will know if it can for three to five years, until big medical studies have been completed. Would this be good enough for you?

Something similar happened during the HIV epidemic in the 1980s. Studying the effect of drugs on the HIV virus meant waiting for a slow and debilitating illness to progress to AIDs. Waiting such a long time for trials meant that more and more people suffered these serious outcomes because drugs could not be authorised speedily enough. Doctors started to consider approving drugs on the basis of treatment effects on earlier measures. They suggested that these earlier measures, surrogates, could be used to predict treatment effect on AIDs progression. Then clinical trials could be shortened and beneficial treatments could be prescribed as fast as possible.

Making sure surrogates can predict treatment effect is not easy, as they can be influenced by many outside factors. Researchers decided that statistical tools for surrogacy evaluation were needed and I have investigated these in my work.

There are many medical areas where illnesses are assessed using categories that are ordered from highest to lowest. These ordered categories do not always have the same increase in between categories. These kinds of measures are known as ordinal. For example, after a patient has an acute stroke the doctor might record a measure of the patient's level of death and disability ranging from no symptoms, minor symptoms up to coma and death. This is known as the Oxford Handicap Scale. You will agree that the difference between no disability and minor symptoms and the difference between coma and death are not the same, but that there is in increase in severity as the scale moves up. Therefore, this is an example of an ordinal measure.

In my work I changed the statistical tools I talked about above so that they could also investigate surrogates that are ordinal. I tested my new statistical tools to make sure that they work in all kinds of settings and found that in general they worked well. Therefore, this work should let doctors working in medical areas where ordinal measures are used to investigate surrogates and hopefully reduce the length of trials on new drugs in these areas.

List of abbreviations:

ACE:	Average causal effect
ACN:	Absolute causal necessity
Binary-ordinal:	The setting where there is a binary surrogate and an ordinal true outcome
CA:	Causal association
CE:	Causal effects paradigm
CEP:	Causal effect predictiveness
Continuous-continuous:	The setting where there is a continuous surrogate and a continuous true outcome
CW:	Christopher Weir
DCE:	Distributional causal effect
DVT:	Deep venous thromboembolism
EP:	Entropy power
HE:	Hannah Ensor
IPC:	Intermittent pneumatic compression
LRF:	Likelihood reduction factor
OHS:	Oxford Handicap Scale
Ordinal-binary:	The setting where there is an ordinal surrogate and a binary true outcome
Ordinal-ordinal:	The setting where there is an ordinal surrogate and an ordinal true outcome
PE:	Pulmonary embolism
PIG	Proportion of information gain
PS:	Principal stratification
PTE:	Proportion treatment effect explained
RL:	Robert Lee

STE:	Surrogate threshold effect
VRF:	Variance reduction factor

List of notation:

S	surrogate
T	true (primary) outcome
Z	treatment
i	trials, where $i = 1,..,N$;
j	patients within a trial, where $j = 1,..,n_i$;
N_T	the total number of patients in all trials, $N_T = \sum_i^k n_i$
U	an unobserved confounding variable
μ_s	intercept values of the model where S is regressed on Z
μ_t	fixed intercept values of the model where T is regressed on Z
α	fixed treatment effect of S
β	fixed treatment effect of T
β_S	fixed treatment effect of T given S
(m_{S_i}, m_{T_i})	random intercept values for joint mixed model of S and T regressed on Z
(a_i, b_i)	random treatment values for joint mixed model of S and T regressed on Z
$(\varepsilon_{S_{ij}}, \varepsilon_{T_{ij}})$	jointly distributed errors for joint mixed model of S and T regressed on Z
$\sum$	variance covariance matric of normally distributed errors $(\varepsilon_{S_{ij}}, \varepsilon_{T_{ij}})$
D	variance covariance matric of normally distributed random effects $(m_{S_i}, m_{T_i}, a_i, b_i)^{\mathrm{T}}$
R^2_{indiv}	measure of individual level surrogacy of the meta analytical approach
R^2_{trial}	measure of trial level surrogacy of the meta analytical approach

$d_{SS}, d_{ST}, d_{Sa}, d_{Sb}, d_{TT}, d_{Ta}, d_{Tb}, d_{aa}, d_{ab}, d_{bb}$ components of D

$\sigma_{SS}, \sigma_{ST}, \sigma_{TT}$ components of $\sum$

θ_p average of canonical correlations over time

R_{Λ}^2 a measure of individual level surrogacy under the meta-analytical approach for repeated measures data

EP() entropy power

R_h^2 individual level surrogacy measure of information theory

R_{ht}^2 trial level surrogacy measure of information theory

ϑ_i family of parameters of a meta-analytic R_h^2

h_i number of deaths in trial i

R_{XOQ}^2 a measure of trail level surrogacy under the information theory approach which accounts for censoring via the Kaplan and Meier (1958) estimator

$\mathrm{s_j}(1), \mathrm{s_j}(2)$ binary surrogate for treatment Z, where Z= 1 or 2, for patient j

$\mathrm{t_j}(1), \mathrm{t_j}(2)$ binary true outcomes for treatment Z, where Z= 1 or 2, for patient j

Y and X random variables

$k_b,\ b \in (1, \dots, m_y)$ values of Y

$k_d, d \in (1, \dots, m_x)$ values of X

p_b probabilities of occurrence of k_b

p_d probabilities of occurrence of k_d

W levels of ordinal true outcome

V levels of ordinal surrogate outcome

$H(\,)$ entropy (for the discrete case)

$h_d(Y)$	differential entropy (for the continuous outcome)
$I(X,Y)$	mutual information
G^2	information gain
LL_0	log-likelihood of generic unsaturated model
LL_1	log-likelihood of generic saturated model
$LL_{T\|S}$	log-likelihood of model of T conditional on Z
$LL_{T\|Z,S}$	log-likelihood of model of T conditional on Z and S
LL_T	log-likelihood of the intercept model of true outcome
p	degrees of freedom
$\gamma_{p:\alpha}$, $\delta_{p:\alpha}$	draws from the non-central χ^2 distribution
n_k	the number of observations the models are based on
$\hat{\varphi}$	maximum likelihood
$\theta_0, \theta_1, \theta_2$	parameters from the model of T regressed on Z and S, where θ_0 is the intercept, θ_1 the treatment and θ_2 the surrogate parameter. Subscripts vary depending on the form of the response variables.
$\gamma_0, \gamma_1, \gamma_2$	parameters from stage two models of R^2_{ht}, γ_0 is the intercept, γ_1 and γ_2 are the parameters for the surrogate intercept and treatment estimate variables. Subscripts vary depending on the form of the response variables.
γ_3	intercept parameter from intercept only model of T
$\bar{\mu}_{S_{V_i}}$	mean intercept value of surrogate regressed on treatment

Chapter 1. Introduction

Clinical trials can be lengthy, large and invasive for patients. As such they are expensive and have ethical implications. A substantial area of investigation has grown out of the desire to reduce the length and size of trials. These aspects of the trial will largely depend on the nature of the primary outcome of interest. For instance, whether the outcomes are rare or occur as a result of a lengthy disease process. Sometimes 'stand ins' or surrogates are used in place of primary outcomes and can provide shorter or smaller trials. For example, if the surrogate occurs at a much earlier time point and can reliably predict the treatment effects on the true (or primary) outcome. In this thesis I adopt the surrogate definition of Temple (1999) who state that a surrogate is:

> "…a laboratory measurement or physical sign that is used in therapeutic trials as a substitute for a clinically meaningful endpoint that is a direct measure of how a patient feels, functions or survives and is expected to predict the effect of the therapy".

Some researchers believe that a surrogate can inform on treatment actions on the true outcome, if there is a correlation between the surrogate and the true outcome. This has been shown not to be the case. Baker and Kramer (2003) showed that even where perfect correlation exists a surrogate cannot necessarily predict a treatment effect on the primary outcome. Wang et al. (2012) produced a thorough investigation of correlation approaches and showed how correlation metrics can have misleading results. They showed that a correlation approach without reference to treatment can be influenced by many different factors and is therefore not reliable. More sophisticated statistical approaches to surrogacy evaluation are required – many have been proposed.

Ordinal outcomes have long been important measures in medical research, for example, the Oxford Handicap Scale (OHS) in stroke (Bamford et al., 1989). These measures are frequently used as primary outcomes in clinical trials. A surrogate outcome methodology that can deal with ordinal outcomes would allow surrogates to be legitimately investigated and adopted in these clinical areas. The current

methodology for surrogate evaluation of ordinal outcomes is limited; that which does exist requires development and refinement.

In this thesis, I outline the work I conducted to extend the current foremost approach to surrogacy evaluation to the case of one or more ordinal outcomes.

In order to achieve this, I conducted a systematic review of all the surrogate evaluation methodology since a previous review by Weir and Walley (2006). My review: investigated the current foremost approaches for surrogacy evaluation; investigated the understanding and perceptions of researchers in this area; and determined the best approach for extension to the case of ordinal outcomes.

The three ordinal outcome settings I was interested in were the case of: a binary surrogate and ordinal true outcome (hereafter referred to as *binary-ordinal*); an ordinal surrogate and binary true outcome (*ordinal-binary*); and an ordinal surrogate and true outcome (*ordinal-ordinal*). I developed the methodology for these three settings. Then, I thoroughly investigated how the methodology worked in real life scenarios using simulation and case studies.

I performed a thorough investigation of the best means of conducting the simulations in the surrogate context. My simulation incorporated a wide range of scenarios that might be expected to occur in real life settings. The simulations were complemented with case studies on the randomised clinical trial CLOTS3 (2013). In these case studies, relevant surrogacy clinical questions of interest were investigated and methodological issues were demonstrated.

These simulation and case studies: determined the usefulness of the developed methodology; provided further information on the benefits of the chosen technique; and informed on issues relating to surrogacy evaluation in the context of ordinal outcomes.

In what follows, the systematic review is discussed in Chapter 2. The methodology, simulation of the: binary-ordinal setting are discussed in Chapter 3 and Chapter 4; ordinal-binary setting in Chapter 5 and Chapter 6; and ordinal-ordinal setting in

Chapter 7 and Chapter 8. The case studies for all settings are presented in Chapter 9. Finally, I conclude in Chapter 10.

Chapter 2. Systematic review

I aimed to update a previous systematic review by Weir and Walley (2006) on the theoretical statistical development of methods for validating surrogates. Molenberghs et al. (2004) and Burzykowski et al. (2005) summarised the early development of surrogacy evaluation. Lassere (2008) provided a useful overarching review of the evolution of the practice of using and appraising surrogates, including early statistical approaches to surrogate outcome evaluation. However, these papers do not discuss all relevant approaches to surrogacy evaluation. This review focuses on statistical methodology developments including the substantial advances that have occurred recently. It also highlights the fundamental differences in the current statistical evaluation frameworks and their advantages/disadvantages. This review will provide a thorough examination of the understanding and perceptions of investigators in this area of research. It will also provide an invaluable resource when attempting to fulfil the aim of extending the current methodologies to ordinal variables.

The search process of this review was conducted in three parts. Firstly, before this PhD was undertaken a systematic literature search was conducted by Robert Lee (RL) from Jan 2003 to 16th of May 2011, after initial screening all papers were reviewed by both RL and Christopher Weir (CW) to determine which contributed to the methodology of surrogate endpoint evaluation. In part two, I updated the review from Jan 2011 to Feb 2013 following the same process as the first stage. I: searched the literature; conducted the initial screening; CW and I then reviewed the remaining papers to determine if they contributed to the methodology of surrogate outcome evaluation. In the third stage, an update of the review was conducted in September 2015 covering the period January 2013 to 15th of September 2015 alongside an updated citation search.

A search of the literature was conducted using MEDLINE and Web of Knowledge, papers were located using the search terms:

Statist* AND (evaluat* or validat*) AND (surrogate OR biomarker)

After the papers were reviewed to determine which contributed to the methodology, a citation search was conducted on papers of specific interest (Buyse and Molenberghs, 1998), (Frangakis and Rubin, 2004), (Joffe and Greene, 2009) and (Alonso and Molenberghs, 2006). This was done via google scholar, MEDLINE and Web of Knowledge. These were then reviewed by one of HE, CW or RL to determine which contribute to the evaluation of surrogacy.

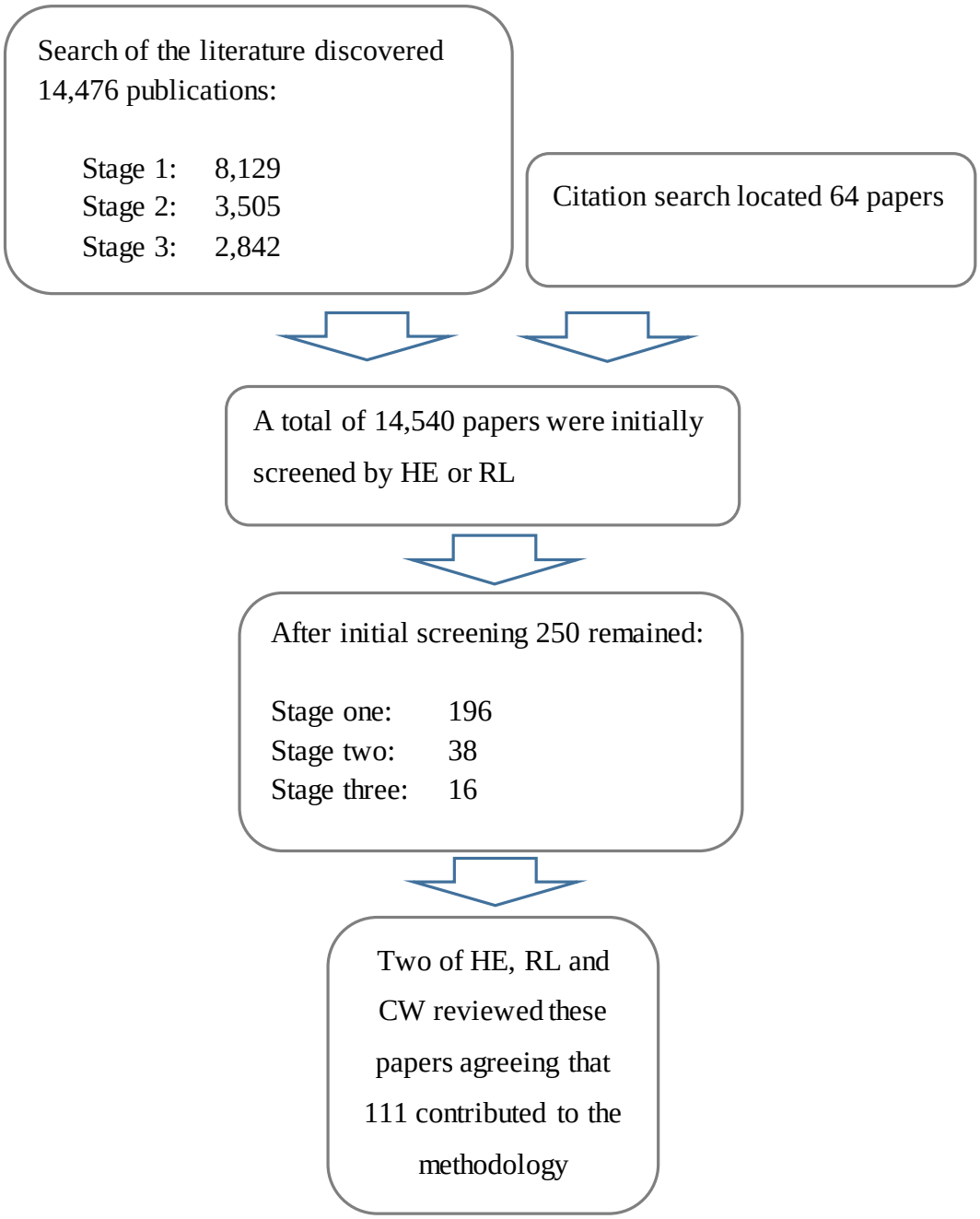


Figure 2.1: *Diagram of review process: incorporating all stages of the review*

Including all three parts of the review a total of 111 papers were deemed to have contributed to the statistical methodology of surrogacy evaluation. The full set of papers from Jan 2003 –Sep 2015 was thoroughly examined by me and the results of this work can be seen in sections 2.1 to 2.9.

The structure of this chapter is as follows. In section 2.1, I describe the seminal papers in this area. I move on to multi-trial approaches in section 2.2: the meta-analytical and information theoretic approaches. Section 2.3 covers approaches dedicated to causal validity: principal stratification and direct and indirect effects. Section 2.4 describes a proposed division of approaches into two paradigms. Sections 2.5 to 2.8 discuss interdisciplinary and miscellaneous approaches, surrogacy schemes (which establish broad evaluations of surrogacy worth) and practical issues. I conclude with a discussion in section 2.9.

In the notation of this review: S and T represent the surrogate and true (primary) outcome respectively; Z represents treatment; i represents trials, where $i = 1, .., N$; j represents patients within a trial, where $j = 1, .., n_i$; $N_T = \sum_i^k n_i$ represents the total number of patients in all trials and U represents an unobserved confounding variable.

Unless otherwise stated: causal validity approaches, in section 2.3, describe methodology based on binary surrogate and true outcomes; all other sections base their methodology on continuous outcomes. Finally, approaches other than those described in the multi-trial section, section 2.2, are based on single trials.

2.1 Landmark and formative proposals

2.1.1 Prentice

Prentice (1989) first recognised the need for a statistical understanding of surrogacy. He defined a surrogate 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 outcome", as well as providing a set of operational criteria for time-to-event outcomes. The 'main' criterion of Prentice is:

$$f(T|S, Z) = f(T|S) \qquad (1)$$

Therefore, T conditional on S is independent of treatment regimen, in other words the surrogate fully describes the effect of treatment on the true outcome. Whilst this publication was ground-breaking in the field and inspired many subsequent developments the criteria have been criticised as being too stringent and of a poor formation for surrogacy evaluation (Freedman et al., 1992), (Fleming et al., 1994).

2.1.2 Proportion of treatment effect explained (PTE)

The main criterion of Prentice is used by Freedman et al. (1992) who developed the proportion of treatment effect explained (PTE) based on binary outcomes.

$$\mathrm{PTE} = 1 - \frac{\beta_S}{\beta} \qquad (2)$$

where β_S and β are the estimates of the effects of Z regressed on T modelled with and without conditioning on S respectively. PTE quantifies the level of surrogacy and can be interpreted as the proportion of treatment effect on the true outcome that is explained by the surrogate. PTE=1 constitutes a perfect surrogate. However, PTE has been criticized as it does not always lie between 0 and 1, therefore it is not a true proportion. Furthermore, it requires that there is no interaction between the surrogate and treatment and encounters issues of imprecision (Weir and Walley, 2006). Several attempts have been made to redefine the PTE in a more meaningful form: Chen et al. (2003a) provided easily computed confidence intervals; Wang and Taylor (2002) proposed an alternative which requires fewer assumptions; Cowles (2002) based their PTE measure on a Bayesian and Huang and Huang (2010) on a counterfactuals approach (a hypothetical manipulation of the experiment, to ascertain the values patients would have experienced had they been on the alternative treatment to that allocated under randomisation).

2.1.3 Relative effect and adjusted association (RE)

Buyse and Molenberghs (1998) proposed another measure derived from the criteria of Prentice known as the relative effect .

$$Relative\ effect = \frac{\beta}{\alpha} \quad (3)$$

Where α and β are estimates the effect of S and T regressed on Z respectively. Relative effect enables prediction of the effect that treatment has upon the true outcome from the information gathered on the surrogate. The relative effect can be thought of as the slope of a regression of the effect of treatment on the true outcome against its effect on the surrogate. Hence relative effect is based on a regression that is assumed accurate based on one data point (the relative effect calculated from one trial only): this is a strong and untestable assumption.

A further measure of surrogacy the adjusted association was proposed in the same paper. The adjusted association measures the association at the individual patient level between the true and surrogate outcome after adjustment for treatment. This paper was amended to correct for minor errors (Buyse et al., 2000a).

2.2 Multi-trial approaches

2.2.1 Accuracy and predictive power – Meta-analytical approach

Following the criticism of RE, that predictive power and accuracy is limited when relying on only one data point, a measure calculated based on multiple REs from multiple trials was proposed by Buyse et al. (2000b). These authors evaluated surrogacy at two levels the individual trial level and the individual patient level; henceforward referred to as the trial level and individual level respectively.

Buyse et al. (2000b) based their trial level measure of surrogacy on the linear regression of the treatment effects of S and T for each trial. If the same relationship between the treatment effects of S and T is seen across trials, S is a good surrogate. (It is also possible to use centres within a trial in place of separate trials if multiple trials are not available; hereafter for simplicity we use the word trial to cover either situation). This is measured via the coefficient of determination (R^2) which captures the strength of the relationship between the treatment effects via the level of variation in the data that is explained by their linear regression. The modelling process for a random effects approach using a joint model is:

$$S_{ij} = \mu_S + m_{S_i} + \alpha Z_{ij} + a_i Z_{ij} + \varepsilon_{S_{ij}} \quad (4)$$

$$T_{ij} = \mu_T + m_{T_i} + \beta Z_{ij} + b_i Z_{ij} + \varepsilon_{T_{ij}}$$

where (μ_s, μ_T) and (α, β) are fixed intercepts and treatment effects respectively. (m_{S_i}, m_{T_i}) and (a_i, b_i) are random intercepts and treatment effects for the i^{th} trial respectively. Error terms in (4) are jointly distributed, $(\varepsilon_{S_{ij}}, \varepsilon_{T_{ij}}) \sim \mathrm{N}(0, \sum)$ and random effects, $\left(m_{S_i}, m_{T_i}, a_i, b_i\right)^{\mathrm{T}} \sim \mathrm{N}(0, \mathrm{D})$.

$$\sum = \begin{pmatrix} \sigma_{SS} & \sigma_{ST} \\ & \sigma_{TT} \end{pmatrix} \quad (5) \qquad \text{and} \qquad D = \begin{pmatrix} d_{SS} & d_{ST} & d_{Sa} & d_{Sb} \\ & d_{TT} & d_{Ta} & d_{Tb} \\ & & d_{aa} & d_{ab} \\ & & & d_{bb} \end{pmatrix} \quad (6)$$

Buyse et al. (2000b) then proposed the use of the coefficient of determination as an expression of the validity of a surrogate at the trial level which they christened R^2_{trial}, see (7). The authors also proposed a measure of surrogacy at the individual level called R^2_{indiv}, see (8). This is the squared correlation between S and T after adjusting for treatment effect and trial. A potential surrogate is said to be perfect at the trial level if $R^2_{\text{trial}} = 1$ and at the individual level if R^2_{indiv}=1.

$$\mathrm{R}^2_{\text{trial}} = \frac{\begin{pmatrix} \mathrm{d_{Sb}} \\ \mathrm{d_{ab}} \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} \mathrm{d_{SS}} & \mathrm{d_{Sa}} \\ \mathrm{d_{Sa}} & \mathrm{d_{aa}} \end{pmatrix}^{-1} \begin{pmatrix} \mathrm{d_{Sb}} \\ \mathrm{d_{ab}} \end{pmatrix}}{\mathrm{d_{bb}}} \quad (7) \qquad \text{and} \qquad R^2_{\text{indiv}} = \frac{\sigma^2_{ST}}{\sigma_{SS}\sigma_{TT}} \quad (8)$$

As an alternative to the random effects approach described above, which can be computationally burdensome, a fixed effects approach can also be conducted using a two stage model. At stage one, the following joint model is fitted:

$$S_{ij} = \mu_{S_i} + \alpha_i Z_{ij} + \varepsilon_{S_{ij}} \quad (9)$$

$$T_{ij} = \mu_{T_i} + \beta_i Z_{ij} + \varepsilon_{T_{ij}} \quad (10)$$

where $(\varepsilon_{S_{ij}}, \varepsilon_{T_{ij}}) \sim \mathrm{N}(0, \sum)$. At stage two, the intercept and treatment estimates of stage one are used as fixed explanatory variables against a response variable indicating which aspect of the model the estimates refer to. The errors of the stage

two model are equivalent to the random effects parameters in (4) again these are distributed $\left(m_{S_i}, m_{T_i}, a_i, b_i\right)^{\mathrm{T}} \sim \mathrm{N}(0, \mathrm{D})$. Hence all the components of the random effects approach can be calculated using a two stage fixed effects approach.

2.2.2 Model fitting improvements– Meta-analytical approach

Many authors have investigated the theoretical, practical and computational issues relating to the meta-analytical approach. For example, there was found to be a significant negative effect on the convergence of models if the number of trials or the between trial variability were small for the joint random effects approach (Buyse et al., 2000b). (Tibaldi et al., 2003) corroborate these findings and recommended using a fixed effects approach, since loss in statistical efficiency is only minor. They suggest that instead of the joint random and joint fixed effects models described in the previous section, (4) or (8) and (9) respectively, one could use the equivalent approaches with individual modelling of treatment on S and on T. At the second stage they suggest modelling:

$$\hat{b}_i = \lambda_0 + \lambda_1 \hat{\mu}_{S_i} + \lambda_2 \hat{\alpha}_i + \varepsilon_i \qquad (11)$$

Where $\hat{\beta}_i$ are the estimated trial specific treatment effects of T, see (10), and $\hat{\mu}_{S_i}$ and $\hat{\alpha}_i$ are the estimated trial specific intercepts and treatment effects of S respectively, see (9). An alternative measure of R^2_{trial} is then calculated from the coefficient of determination of (11). (Note: the use of individual models mean that $\sum$ and hence $\mathrm{R}^2_{\text{indiv}}$ are not easily determined). Abrahantes et al. (2004) considered hierarchies within trials, for instance where individuals are modelled within centres within trials, and the influence of ignoring a hierarchical level. They concluded that, if the level of surrogacy was the same at trial and centre level a model which ignored a hierarchical level performed well. If the levels were different ignoring the trial level can lead to overestimation of the variability and biased estimates of the centre level association. However, trial and centre level results were similar if variation at the centre level was lower than that at the trial level which may justify the use of centre instead of trial in practice. Abrahantes et al. (2008) noted that the meta-analytical approach has an enforced form and an implicit assumption that all the relationships are linear, and they suggested alternatives. They also suggested that

cross validation of the meta-analytical approach would prevent overly optimistic estimates. Further computational issues were investigated in Tilahun et al. (2007) in regard to the coding of treatment and ill conditioned matrices.

2.2.3 Extensions to alternative settings- Meta-analytical approach

We now focus on the extension of Buyse et al. (2000b) to the setting where the true or surrogate outcomes are time-to-event, binary, ordinal or repeated measures. Authors reformatted the models of the meta-analytical approach to fit the data types required, for example Cox models or generalised linear models, in order to extend the approach to alternative settings. In each setting it is possible to calculate R^2_{trial} via the coefficient of determination exactly as in the continuous case; hence this process will not be re-described in the following sections. However, in order to calculate the correlation coefficients used to calculate R^2_{indiv}, see (5) and (8), joint modelling is required and various complications arise when this is conducted in settings with non-continuous outcomes. Alternative measures of R^2_{indiv} are therefore required and those proposed are described alongside other relevant methodology. Due to the non-uniform measurement of R^2_{indiv} across settings Dai and Hughes (2012) proposed an alternative approach to the meta-analytical approach using estimating equations which is more readily extendable to other settings.

2.2.3.1 Time-to-event

Burzykowski et al. (2001) extended the meta-analytical approach for outcomes that are time-to-event variables. They proposed the use of a copula to model the joint survival function of (S_{ij}, T_{ij}) since other options are less flexible. In this setting, non-linear relationships are more likely because the correlation across trials is assumed not to be consistent, making the use of the correlation coefficient for R^2_{indiv} unviable. The authors therefore proposed the use of Kendall's τ. After experiencing convergence problems with this approach, Renfro et al. (2012) substituted the second stage of modelling with a Bayesian alternative which they found avoided the undesirable assumptions of no measurement error and common baseline hazards across trials. As a further alternative, Tibaldi et al. (2004) suggested the use of the Plackett-Dale model (Plackett, 1965) for multivariate data, which can measure

dependence between two time-to-event outcomes and can be summarized via an odds ratio θ, Kendall's τ or Spearman's ρ. Burzykowski et al. (2003) considered the case where the surrogate is ordinal and the true outcome is a time-to-event variable. They used the same approach as Tibaldi et al. (2004) except they modelled the surrogate outcome in (9) with a proportional odds model where $\hat{S}$ is a latent continuous variable.

An approach to adjusting for semi-competing risks, encountered in the meta analytical approach if S experiences dependent censoring via T, was described for the single trial measures in Ghosh (2008b) and Ghosh (2009) and extended for the meta-analytical approach in Ghosh et al. (2012a). They used the accelerated failure time model to account for and estimate dependent censoring, and dissociated the disease process and true outcome by investigating how the disease would have behaved had T not occured. They analysed S as a latent variable, which was then constrained to the region S≤T, called the region constraint, and made use of an artificial censoring technique to account for dependent censoring. There is some debate surrounding the authors use of the region constraint: Molenberghs (2012) described this approach as "elegant". Berger et al. (2012) argued that region constraint does not describe reality and preferred a composite approach, incorporating the outcome and the censoring mechanism. Ghosh et al. (2012b) contended that the composite approach is not appropriate since the surrogate incorporates information on the true outcome.

Other useful contributions were made by Abrahantes and Burzykowski (2010) who applied to time-to-event outcomes the simplified modelling techniques suggested in Tibaldi et al. (2003). However, these strategies were not found to be appropriate in this setting as considerable levels of bias occurred in the calculation of R^2_{trial}. This was thought to occur because their simplified approach ignored individual level association and ignoring a hierarchal level can bias results, as previously discussed (Abrahantes et al., 2004). Renfro et al. (2014) perform a comprehensive assessment of the bias due to using centres instead of trials for surrogacy assessment in a wide range of scenarios for time-to-event outcomes. They found the conclusions based on using centre would be similar to when trial had been used, however bias using centre

was larger in a number of scenarios. Shi et al. (2011) used a simulation study to compare conventional (non-joint model) and model based approaches (using joint models) at the trial level. They found that the conventional approaches have similar performance to model based approaches but with fewer computational difficulties.

2.2.3.2 Binary and ordinal

Molenberghs et al. (2001) extended the meta-analytical approach to the case where one outcome is binary and one continuous. They used a latent variable to represent the continuum underlying the dichotomized variable and applied a generalized linear model leading to an R^2_{indiv} of ρ^2 calculated from the covariance matrix. Renard et al. (2002) used continuous latent variables represented by two observed binary surrogate and true outcomes. They then applied a multilevel probit model from which they identified parameters of interest using pseudo-likelihood techniques. In the case of ordinal outcomes, Alonso et al. (2002) investigated application of this methodology to psychiatry where surrogate and true outcomes could arguably be interchangeable.

2.2.3.3 Repeated measures

Alonso et al. (2003) proposed using time specific functions in (9) and (10) to take a repeated measurement surrogate and true outcome into account. They noted that the assumption in Buyse et al. (2000b), that the error variance covariance matrix $\sum$ in (5), is constant over all trials, is not appropriate in this setting. This is because repeated measures could differ between trials both in number and collection time. Therefore they suggested using the variance reduction factor (VRF), in place of R^2_{indiv}, to allow covariance structures to vary across trials. The VRF is calculated by summing, within each trial and then over all trials, the trace of the variance covariance matrix for repeated measurements on the true outcome and the true outcome given the surrogate. This was used to calculate the reduction in variance on the repeated measurements of the true outcome that can be attributed to incorporation of the surrogate. Alonso et al. (2004a) developed an individual level surrogacy evaluation technique based on a canonical correlation interpretation of the VRF. Canonical correlation identifies variables that explain shared variance through investigation of the variance covariance matrix of two different variables. The authors identified the measure θ_p, related to the canonical correlations as a means of

evaluating surrogacy at the individual level for repeated measures, which can be thought of as the average of the canonical correlations over all trials. Though the authors pointed out that θ_p is symmetric and invariant for linear bijective transformations, it relies heavily on the normality assumption. As an alternative, (Alonso et al., 2006) introduced R_Λ^2 which is both symmetric and invariant and is also more readily extendable to non-normal settings. As with the VRF, a R_Λ^2 close to one indicates a good surrogate. They then showed that θ_p can approximate R_Λ^2.

Pryseley et al. (2010) proposed a means of calculating the optimum number of repeated measurements of the surrogate in terms of the cost, using a discrete true outcome. Weights are added to incorporate a researcher's view of the importance of the cost of collecting repeated measurements versus increased precision in evaluation through collecting more repeated surrogate measures.

Renard et al. (2003) investigated a repeated measures surrogate outcome for a time-to-event true outcome. They used the simplified model (11) to calculate R_{trial}^2 and the joint model using a latent Gaussian process for R_{indiv}^2. Tilahun et al. (2009) investigated the case where one outcome is a repeated measure and the other is a cross sectional variable. The measures they propose are variations of the VRF and R_Λ^2 with different results depending on which outcome is cross sectional. They found that their VRF alternative predicts the repeated measures sequence as a whole when the surrogate is cross sectional, whereas their R_Λ^2 predicts an 'optimal linear combination' of the repeated measures.

2.2.4 Clinical interpretation– Surrogate threshold effect

Another consideration in the meta-analytical approach is that there is no c ϵ [0,1] such that $R_{trial}^2 \geq c$ constitutes a good surrogate. Any such value would be completely arbitrary; therefore it is difficult to interpret the practical worth of any R_{trial}^2 value. The surrogate threshold effect (STE) based on the meta-analytical approach, was proposed by Burzykowski and Buyse (2006) to aid interpretation. The STE is defined as "the minimum value of a treatment effect on a surrogate outcome which the predicted effect on the true outcome would be significantly different from zero." The larger the variance the larger the absolute value of STE, so the STE can

be thought of as a measure of the precision of the prediction of treatment effect on the true outcome. This is useful in practice as a large STE would indicate a large treatment effect on the surrogate is required before the true unknown treatment effect on the true outcome can be identified reliably from this data alone, suggesting a poor surrogate. Burzykowski and Buyse (2006) stated that a further benefit of STE is that it quantifies the loss of efficiency in estimating the treatment effect on the true outcome using the surrogate as opposed to the true outcome. One can then determine if this loss of efficiency is acceptable taking into account the corresponding reduction in trial duration. Johnson et al. (2009) assessed the STE in practice, using trial level data, and concluded that it is a straightforward promising measure of surrogacy. They noted that prediction intervals for the STE vary according to the size of a future trial being predicted, larger trials provide narrower confidence intervals.

2.2.5 Unification – Likelihood Reduction Factor

As seen in section 2.2.3 when the meta-analysis approach is extended to non-continuous outcomes the interpretation of R^2_{indiv} is incoherent across settings: measures take different forms and are sometimes assessed at a latent level. The likelihood reduction factor (LRF) is a quantification of individual level surrogacy that is applicable regardless of the type of outcome studied (Alonso et al., 2004b). This is based on the amount of information gained about the true outcome after accounting for the surrogate. These authors proposed the use of the generalised linear model versions of (10) and (12) for each trial i, for binary outcomes, which regress the true outcome on treatment with and without adjustment for the surrogate respectively:

$$T_{ij} = \theta_{0_i} + \theta_{1_i} Z_{ij} + \theta_{2_i} S_{ij} + \varepsilon_{T|S_{ij}} \qquad (12)$$

The difference in the amount of information on the true outcome gained from the surrogate is calculated via the difference in the log-likelihood between (10) and (12) which is formally expressed as G_i^2, for each trial i. L_0 is always the log-likelihood for the unsaturated model, in this case (10), and LL_1 for the saturated model, (12), for trial i. $G_i^2 = 2 * (LL_1 - LL_0)$. The LRF is then calculated:

$$LRF = 1 - \frac{1}{N}\sum_{i} \exp\left(-\frac{G_i^2}{n_i}\right) \quad (13)$$

The authors stated that the LRF reduces to the R^2_{indiv} measure where outcomes are normally distributed.

Qu and Case (2007) proposed an alternative measure: the proportion of information gain (PIG). Li and Qu (2010) investigated the PIG in relation to measurement error. Alonso et al. (2007) questioned the conceptual basis of PIG and listed several drawbacks. Miao et al. (2012) proposed non parametric versions of PIG and the LRF which do not assume a pre-specified form and performed better than the original versions under simulation study. Using Prentice's main criterion the authors also proposed a safety measure which may help avoid potential type I errors.

One criticism of the LRF is that, like PTE, it cannot account for the presence of an interaction between the surrogate and treatment in (12). A further critique, that it has no population level interpretation has been resolved through the development of an information theory approach.

2.2.6 Population level interpretation- Information theory approach

Information theory is chiefly concerned with the quantification of the level of uncertainty in a random variable. Entropy is a fundamental concept which is related to how readily (or with how much certainty) one can guess an observed value of a random variable. If an event is expected, then its occurrence does not provide as much information (has lower entropy) as if the converse were true. An extension of this concept proposed by Shannon and Weaver (1948) is a measure called entropy power (EP) which can be used to compare continuous random variables.

A convenient aspect of information theory is that one can quantify the amount of uncertainty in a variable expected to be reduced if information about another variable is known. This is called the mutual information. In the case of surrogate outcomes, if the surrogate explains a lot of uncertainty surrounding the true outcome then the mutual information will calculate the corresponding reduction in its entropy. This

quantity can be considered as the information in T that is shared by S. Alonso and Molenberghs (2007) used these concepts to suggest an information theory measure of surrogate association.

$$R_h^2 = \frac{\mathrm{EP}(T) - \mathrm{EP}(T|S)}{\mathrm{EP}(T)} \tag{14}$$

where $EP(T)$ is the entropy power of T and $EP(T|S)$ is the entropy power of T given S. This can be thought of as the proportion of uncertainty in T at the individual level removed by adjusting for S. R_h^2 can also be mathematically linked to the mutual information. These concepts are in line with the aims of surrogate evaluation since surrogacy is concerned with increasing our knowledge about treatment effect on the true outcome through the use of the surrogate. Hence $R_h^2 \approx 1$ indicates a good surrogate that removes almost all uncertainty in the true outcome. R_h^2 is consistent across different settings, providing a unified approach. Alonso and Molenberghs (2007) also proposed a version of R_h^2 applied at the trial level using the full hierarchical approach from Buyse and Molenberghs (1998). This they called the R_{ht}^2:

$$R_{ht}^2 = 1 - \frac{\mathrm{EP}(\beta_i|\alpha_i)}{\mathrm{EP}(\beta_i)} \tag{15}$$

where α_i and β_i are the treatment effects on the surrogate and true outcome respectively. And can be interpreted as the proportion of uncertainty in the treatment effects on T removed by adjusting for treatment effects on S. R_{ht}^2 can be shown to reduce to R_{trial}^2 in the bivariate continuous setting.

In order to account for heterogeneity among trials Alonso and Molenberghs (2007) proposed a meta-analytic R_h^2, where N trials produce N_q possible $R_{h_i}^2$:

$$R_h^2 = \sum_{i=1}^{N_q} \alpha_i R_{h_i}^2 \text{ where } \vartheta_i > 0 \; \forall \; i \text{ and } \sum_{i=1}^{N_q} \vartheta_i = 1 \tag{16}$$

They note that the choice of ϑ_i represents an uncountable family of parameters. However they highlight that the LRF is a family member which supports unification through its common interpretation across settings and is a consistent estimator of R_h^2. This gives the LRF a connection to the population measure of surrogacy which was

previously unrecognised. The LRF is therefore a very useful and appropriate option for a multi-trial measure of R_h^2.

Information theory provides a much needed unified approach. Alonso and Molenberghs (2008) noted additional advantages compared with the meta-analytical approach which include; narrower confidence intervals; joint models (which previously caused computational issues) are not used, and latent variables (which hamper interpretation) are not required.

2.2.7 Extensions to alternative settings- Information theory approach

All extensions of information theory approach to different settings use the LRF to calculate R_h^2. As with the meta-analytical approach the models of the LRF, (10) and (12), are reformatted to fit particular data types. All published extension papers have been able to apply the LRF via (13) and derive R_h^2 in a format consistent with other settings.

2.2.7.1 Time-to-event outcomes

A time-to-event information theory approach published by Alonso and Molenberghs (2008) suggested that, at the trial level, if trial specific treatment effects on S and T are linear then R_h^2=R_{trial}^2, where R_{trial}^2 is calculated according to Burzykowski et al. (2001), otherwise R_h^2 is the better measure of surrogacy. At the individual level, the authors fitted two survival models for (10) and (12) and then applied the LRF via (13). Pryseley et al. (2011) investigated information theory extensions to the case where censoring occurs, using the measures of O'Quigley and Flandre (2006). In the information theory setting Pryseley et al. (2011) named these measures LRF-a and R_{XOQ}^2. LRF-a is weighted by h_i, the number of deaths in each trial i, and partly accounts for the censoring mechanism. R_{XOQ}^2 is a LRF measure using the Kaplan and Meier (1958) estimator, chosen because it consistently estimates the distribution function in the presence of censoring and is not impacted if censoring is independent (O'Quigley and Flandre, 2006). Pryseley et al. (2011) showed that in the information theory case the LRF-a did not account for censoring well but as long as the amount of semi-competing risks was not high R_{XOQ}^2 was a good estimator of R_{h}^2.

2.2.7.2 Binary outcomes

Pryseley et al. (2007) extended the information theory approach for one binary and one continuous outcome. They used generalised linear models, in place of (10) and (12), to calculate the LRF. They found that in the cross sectional case the LRF reduces to R^2_{indiv}. Tilahun et al. (2008) compared information theory measures with the meta-analysis and simplified meta-analysis approaches advocated by Tibaldi et al. (2003) for the binary-binary case. They found the information theory approach superior, since it encountered fewer issues with model convergence and was easier to interpret (due to lack of latent variables). They also reported that simplification techniques worked well as long as data are available from many large trials.

2.2.7.3 Repeated measures

Alonso et al. (2006) included functions of time into the models used to determine the LRF for the repeated measures setting. Where both surrogate and true outcomes are repeated measures, the LRF is equivalent to R^2_Λ, see section 2.2.3.3. Where one is a repeated measure and the other a cross sectional variable, the LRF reduces to R^2_{indiv}.

2.3 Causal evaluation

2.3.1 Causal validity- Principal stratification

Randomisation provides intervention groups that are on average balanced in terms of potential confounding variables; therefore, treatment can be reasonably assumed to be the only factor influencing outcome and a causal relationship can be inferred. However, if analysis is adjusted for variables measured after treatment allocation (variables that are not randomised) it becomes possible for comparison groups to be unbalanced in terms of possible confounding factors. In this case, any observed relationship between intervention and outcome may be due to the influence of the confounding factors and therefore outcome cannot be deemed to be caused by intervention. This is called 'post-treatment selection bias'. This bias limits the ability of previously described approaches, in sections 2.1 and the individual level measures of section 2.2, to offer a causal interpretation of surrogacy, since surrogates are adjusted for in analysis of the true outcome and are measured post-randomisation. For a binary surrogate and time-to-event true outcome Frangakis and Rubin (2004) showed that since these approaches ignore the issues surrounding causality it is

possible to validate surrogates where treatment influences the true outcome without affecting the surrogate. They therefore proposed a causal surrogate evaluation approach via principal stratification. Nevertheless, trial level surrogacy under the meta-analytical and information theory approaches (section 2.2) does have a causal interpretation, since causal treatment effects on the true outcome are regressed on the treatment effects on the surrogate: hence the analysis of the true outcome is not affected by a post-treatment variable.

As previously mentioned, in this and the following section, the methodology discussed is based on binary outcomes, unless otherwise stated. Consider a set of patients j=1,…,J; let s(1), s(2), t(1) and t(2) be the binary surrogate and true outcomes for treatment Z, where Z= 1 or 2. In essence principal stratification centres on performing a within patient treatment comparison where every patient returns $s_j(1), s_j(2), t_j(1)$ and $t_j(2)$. Treatment effects on the true outcomes are assessed within basic principal strata which are based on the outcomes of the surrogate i.e. the ordered pair $(s_j(1), s_j(2))$ for each patient. The strata do not vary according to treatment hence there is no selection bias within stratum imposed by the surrogate. Furthermore, any treatment comparison within the strata is based on an identical set of patients and can be said to be causal.

Frangakis and Rubin (2004) proposed a new definition of surrogacy where S is a *principal surrogate* in the case of a comparison between two treatments Z=1 and 2 if, for all fixed *s*, the comparison between ordered sets:

$$\{T_j(1): S_j(1) = S_j(2) = s\} \text{ and } \{T_j(2): S_j(1) = S_j(2) = s\} \qquad (17)$$

results in equality. This definition says that S is a principal surrogate if causal effects of Z on T only occur in the strata where there are causal effects of Z on S. This is because, patients in (17) that have the same true outcome regardless of treatment only occur in the strata where patients have the same surrogate outcome regardless of treatment. It follows that causal treatment effects (where a patient has different outcomes for the two treatments) on the true outcome are restricted to strata that have causal treatment effects on the surrogate.

The authors suggested comparisons across ordered sets to quantify the level of surrogacy in practice. For example, when examining the set of patients where the treatment has a causal effect on the surrogate we consider the proportion of patients who also experience causal effects on the true outcome; this is the *associative* effect. Conversely the proportion with no treatment effect on the surrogate outcome based on the set where there is a causal effect on the true outcome is *dissociative*.

2.3.1.1 Identification- Principal stratification

Principal stratification requires both treatment outcomes to be known but generally only one is observed; in order to estimate unobserved outcomes a counterfactuals approach is adopted. Here, multiple imputation under the assumption of missing or incomplete data may be used to estimate missing responses. In this case, since only one of the treatment effects is observed, the counterfactual model is over parameterised; which leads to issues of identifiability. Additional assumptions are required to aid estimation (Rubin, 2004). Wolfson and Gilbert (2010) highlighted the difficulty of identifying all parameters and investigated a number of assumptions, some more stringent and testable than others, and their influence on the identifiability of principal surrogate measures. Considering the special case of vaccine trials, they found that estimating treatment effects on the true outcome within principal strata is possible but only in the case of certain conceivably improbable assumptions. They suggested a sensitivity analysis as a remedy but conceded that it may be more useful to focus on non-causal measures of surrogacy that are not restricted to principal strata to estimate treatment effects and that can be identified under less stringent assumptions. The strong assumption of monotonicity which might be true for most but not necessarily all patients is investigated by Li et al. (2011). This assumption states that if one treatment is better overall it is not possible for any patient on this treatment to fare worse than on placebo. They found that implementation of the monotonicity assumption could cause extreme bias if even a small number of patients' outcomes violated the rule, which is likely in practice. On the other hand, relaxation of the assumption causes large loss of efficiency causing models to be unreliable.

2.3.1.2 Practical implementation- Principal stratification

Regardless of these issues, there has been much work on developing the practical application of principal stratification. Conlon et al. (2013) extended the approach in Frangakis and Rubin (2004) to the multivariate normal situation. Li et al. (2010) and Li et al. (2011), used a Bayesian imputation technique for the single and multi-trial setting respectively, proposed the *common associative proportion*, which is the proportion of patients that have causal treatment effects on both the surrogate and true outcome as opposed to the proportion that have causal effects on only one. Elliot et al. (2013) extend this approach to investigate a surrogate where there is missing data on the true outcome. The *causal effect predictiveness* (CEP) surface plot provides a quantification of principal surrogacy (Gilbert and Hudgens, 2006) & (Gilbert and Hudgens, 2008). The CEP incorporates a *necessity* requirement (ACN) based on the need for casual effects on the surrogate if witnessed on the true outcome and a *sufficiency* requirement (ACS) similar in spirit to the STE (see section 2.2.4). Functions of the CEP are used to summarise the magnitude of associative and dissociative proportions; the *expected dissociative effect* the *proportion associative effect* and the *associative span*. A further development in a similar vain to the CEP has also been presented for time-to-event clinical outcomes (Qin et al., 2008). Here a measure named *predictive surrogacy* informs a researcher of both the population level properties of necessity and sufficiency. Zigler and Belin (2012) estimated the CEP surface using a Bayesian approach which suffers lower levels of bias through avoidance of untestable assumptions. Their approach incorporates prior information on observed and unobserved outcomes and in certain situations a sensitivity parameter that incorporates unobservable associations. Huang and Gilbert (2011) developed a graphical aid the *standardised total gain* and summary measure *total gain*, to compare the worth of different composite surrogates. Finally, Gilbert et al. (2003) suggested that a sensitivity analysis may be required in the case of post treatment selection bias due to factors other than the surrogate.

2.3.1.2.1 Augmented trial designs: adding practical implementation and identifiability

Follmann (2006) investigated whether immune response variables had a causal role in the risk of outcome of interest. He suggested using augmented trial information to

identify the value of a surrogate (immune response) that a person might have had, had they been randomised to the unobserved treatment (vaccine). He suggested two ways to achieve this, a baseline immunogenicity predictor and a closeout placebo vaccination approach.

The baseline immunogenicity predictors approach uses a baseline variable that correlates highly with the immune response. Based on the baseline variable and observed outcomes they imputed the unobserved immune responses of patients on the placebo.

The closeout placebo vaccination approach is based on providing the vaccine to the placebo group after the trial has completed and then observing their immune response. These post trial immune responses are then used as the unobserved immune responses of the placebo group for the original trial. This assumes that the timing of the application of the vaccine has no impact on outcome. In the context of vaccine trials these approaches aid identifiability and allow researchers to investigate whether immune response is causally related to risk of the outcome of interest (Follmann, 2006).

Using the baseline immunogenicity predictor approach Gabriel and Gilbert (2014) proposed a time and surrogate dependent efficacy curve based on a Weibull model to assess principal surrogacy for time-to-event outcomes with right censoring. Gabriel et al. (2015) extend this approach to assess combinations of biomarkers as surrogates. Huang et al. (2013) used a pseudo score estimator to best utilise augmented data for the estimation of counterfactual results. Liu et al. (2014) suggested that a semiparametric likelihood or pseudo-likelihood (PL) approach is better than the pseudo score estimator and performed a simulation for comparison. They found that the semiparametric likelihood method is superior to pseudo score estimator since it does not require a model for missingness, but that the pseudo-likelihood approach has a computational advantage over the semiparametric likelihood approach and pseudo score estimator.

Miao et al. (2013) investigated these augmented trial approaches and showed how they aided identification. They also investigated whether misclassifying patients into

principal strata using augmented data would bias results of the absolute causal necessity (ACN) introduced in section 2.3.1.2. They found that in the case of "constant biomarkers" in vaccine trials, where placebo patients have no immune response, overall bias imposed by misclassification was small. For the case of "arbitrary" variability there was some bias witnessed, although knowledge of the behaviour of this bias may help interpretation of results.

In light of these and previous developments Wolfson and Henn (2014) examined the principal surrogacy framework and highlighted conditions for partial or full identifiability of the ACN and ACS (described in section 2.3.1.2., which are based on principal strata).

They find that: the stable unit treatment value assumption and ignorable treatment assignment do not provide much identifiability; the assumption that the surrogate is defined at the same time for all patients provides identifiability of some components of ACN and ACS; and the "constant biomarker" assumption for vaccine trials allows identifiability of one of the principal strata required to calculate ACN and ACS. The monotonic assumption can aid identification in some very particular circumstances.

Overall, they show that the augmented trial design approach can allow for full identification of ACN and ACS. However, they do note that certain assumptions are only valid for particular trial types and are unlikely to hold for others. For instance, the constant biomarker and monotonic assumptions are more relevant for vaccine trials and in general will not be suitable. The same principle applies for the augmented trial designs. This thorough examination of identifiability shows that assessing surrogacy using PS can work under some very particular circumstances. However, for more general randomised trials identifiability of the principal stratification approach is still a very large and unresolved issue which hinders the assessment of surrogates in practice.

2.3.1.3 Theoretical issues- Principal stratification

Several authors Pearl (2011), Mealli and Mattei (2012), Pearl and Bareinboim (2011) and Joffe (2011) criticised the theoretical basis of principal stratification with respect to surrogacy suggesting analysis where treatment effects can only be assessed

in separate restricted strata may hinder and limit investigations. Furthermore, they claimed that the approach is not ideally framed to inspect the predictive ability or transportability (in this case, meaning the ability to transfer information learned in one trial to future trials) of a surrogate. Pearl and Bareinboim (2011) showed that since principal stratification is based on only one trial a principal surrogate may not be able to predict treatment effects in a new trial and a good predictor may not constitute a principal surrogate. They suggested a new causal conceptual approach for transportability based on graphical representations but this is of limited use since it cannot be applied in practice (Joffe, 2011). Gilbert et al. (2011) disputed these criticisms, outlining a case where research aims can be specifically addressed using principal stratification. They also suggested a measure investigated through sensitivity analysis or comparison of true and predicted results from multiple trials to tackle transportability. Furthermore, Baker (2006) and Baker (2008) proposed an *average prediction error* which assesses the average absolute difference between the predicted effect of Z on T, via the effect of Z on S, versus the true observed effect for each trial, and an approach based on hypothesis testing. Baker (2008) was amended in Baker (2009) to correct for minor errors.

Chen et al. (2007) identified a paradox in principal stratification where causal treatment effects are seen on the surrogate and true outcome given the surrogate but not the true outcome; this may occur because an unobserved confounding variable, U, acts on both the surrogate and true outcomes. This situation is a matter of some concern in surrogate evaluation literature (Lauritzen, 2004). If this paradoxical situation cannot be identified it would result in the authorisation of treatments that have a negative impact on the true outcome of interest, with potentially fatal consequences.

Principal stratification is susceptible to the so called *surrogate paradox* because it does not consider treatment directions. To avoid this issue the authors propose additional conditions based on treatment direction using the *average causal effect* (ACE). Ju and Geng (2010) showed that in certain situations the ACE can fail to detect the surrogate paradox, and hence they redefined the conditions in terms of a proposed *distributional causal effect* (DCE). Both approaches require knowledge of

U which is unlikely in practice. Kuroki (2013) noted that neither the ACE or DCE gave a calculation of the size of the causal treatment effect on the true endpoint. They proposed the use of sharp bounds to derive 'closed form formulas for upper and lower bounds on the causal effects of Z on Y [the true outcome]'. In other words, they set a range in which the casual effects must lie for a surrogate to be deemed valid. Wu et al. (2011) suggested the ACE and DCE rely on the potentially invalid monotonicity assumption and do not account for equivalence relationships. They proposed conditions that a surrogate must meet in the counterfactual situation to satisfy the original criteria of Prentice, incorporating treatment directions. Gilbert et al. (2015) were also interested in PS, its connections to the surrogate paradox and the ideas of Prentice. Using the measures ACN and ACS of principal stratification they found no direct links to the criteria of Prentice or to the surrogate paradox. However, there was an implied relationship under certain conditions.

2.3.1.4 Principal stratification: discussion

In summary, principal stratification is a valuable approach to surrogate endpoint evaluation chiefly because it permits investigation of causal effects. As Wolfson and Gilbert (2010) stated "potential outcomes provide a natural way of formulating and answering important scientific questions that are difficult to answer without counterfactuals". However, potential outcomes encountered serious issues: stringent assumptions affect the ability to estimate effectively, as a consequence of guarding against the alternative issue of non-identifiability. Identifiability has been found to be possible in some very particular circumstances but in the general case is still intractable. Additional issues surrounding transportability and the surrogate paradox do raise questions about the practical worth of such an approach. Although the issues of transportability and the surrogate paradox have been shown to be resolvable, such resolutions are currently not a central component of the principal stratification approach.

Principal stratification is a promising approach to surrogacy evaluation, which may take a more prominent role in the future. However, given its limitations it is unsurprising that a well-established practical means of estimating surrogacy via principal stratification has yet to emerge.

2.3.2 Alternative causal approach - Direct and indirect effects

Another causal surrogacy evaluation approach concerns the measurement of direct and indirect effects. Consider Figure *2.2* direct effects of the treatment are those that act only on the true outcome and not the surrogate. Indirect effects are those that act through the surrogate. Under a naïve model, as in Figure *2.2*,A.)direct and indirect effects can be determined by empirical evidence observing the proportion of patients that had agreement in treatment effects between the true and surrogate outcomes (indirect effects) and the proportion where treatment only affects the true outcome (direct effects). However if one carries out a such a naïve analysis when more complex causal relationships exist the results may be biased (Emsley et al., 2010). Cox and Wermuth (2004) expored causal models that might occur in more detail, and recommend a sensitivity analysis in the case of unobserved confounding variables, see Figure *2.2*, B.).

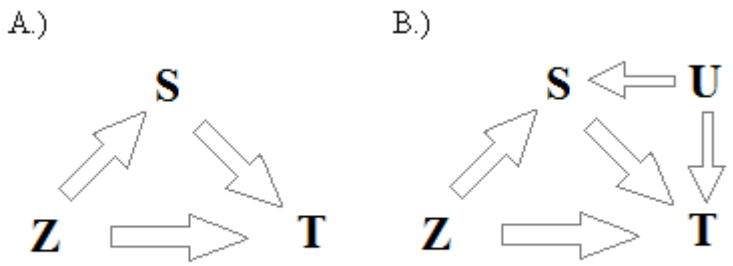


Figure *2.2: Direct and indirect effects of: A.) a naïve model B.) a model where a confounder U acts on S and T.*

Robins and Greenland (1992) investigated the means of identifying direct and indirect effects under a number of causal models. They used a potential outcomes approach under which all the possible outcomes that may have occurred had the treatment been different are investigated. A theoretical example showed that the typical means of measuring direct and indirect effects by adjusting for the surrogate may be biased because there is a striking difference in the true underlying relationships and the conclusions that can be drawn from the observable results. They went on to show that identifying direct and indirect effects is an intractable problem in a number of settings, even in the case of a cross over trial with no carry over.

However, when the surrogate outcome can be controlled, by some randomisation intervention or using information on additional confounding variables, direct and indirect effects can be identified via G-computation.

Qu and Case (2006) used path analysis, in the case where multiple surrogate markers are being assessed, to determine the proportion of direct and indirect effects attributable to each of a collection of surrogate markers. Dibaj et al. (2010) highlighted some errors in the final models but Qu and Case (2010) pointed out that any direct and indirect effects calculated would be valid despite these. Ditlevsen et al. (2005a) investigate mediation variables (where a biological markers lies at least in part in the causal pathway from treatment to outcome), and used empirical evidence to calculate a measure based on the proportion of indirect effects which controls for potential confounders. Kaufman et al. (2005) criticised Ditlevsen et al. (2005a) for using a latent set up and for not adopting a causal approach. Ditlevsen et al. (2005b) further discuss this work in reference to MacKinnon et al. (2007).

Taylor et al. (2005) showed that the alternative PTE measure, PE, of Wang and Taylor (2002) can be expressed in terms of direct and indirect effects. Using assumptions and constraints they found there are some relations, but no strong link, to measures from direct and indirect effects and principal stratification approaches. They concluded that PE is the only measure that can be estimated from the data and that, under untestable assumptions, it may have an 'approximate causal interpretation'.

2.4 Classification of approaches

Joffe and Greene (2009) proposed classification of approaches under two paradigms: the causal effects and causal association paradigm. They considered the original definition of Prentice, the meta-analytical, principal stratification and the direct and indirect effects approaches.

The causal effects paradigm (CE) requires the surrogate to lie in the causal pathway from treatment to the true outcome. This requirement is understandable and desirable but causes problems from a practical point of view. Measuring the association between Z, S and T is straightforward in the naïve scenario but not in a more

complicated situation where an unobserved confounder *U* influences S and T, see Figure *2.2*, B.). In an investigation where *U* exists but is unaccounted for as there is an association of *U*→S and *U*→ T there would appear to be an association of Z→*U* therefore Z→T even if there is no direct effect present. Adjustment for the confounder *U* can be made via certain strong assumptions and where the surrogate can be manipulated, which is not generally practically feasible. Prentice and direct and indirect effects approaches fall under this paradigm.

The causal association (CA) paradigm avoids this issue by not investigating the effect of the surrogate on the true outcome just the treatment effect on the surrogate and true outcome separately as in both the meta-analytical and the principal stratification approaches. There is no direct measurement of the relationship of the true outcome and surrogate calculated; therefore, issues related to modelling of the causal pathway are not relevant.

The authors indicated that there are some connections between the two paradigms and situations in which models in each paradigm are related but interpretation of models and parameters are not the same. They went on to state that the causal association paradigm is more useful in practice because it does not suffer issues of bias in the presence of confounding or interaction between variables. They also gave strong arguments for a multi-trial approach to surrogacy evaluation such as can be found in the meta-analytical and information theory approaches. However, they expressed regret that the current multi-trial approaches do not have a strong causal basis.

2.4.1 Relationships within classifications of Joffe and Greene (2009)

Alonso et al. (2014) investigated the relationship between the meta-analytical and the average causal necessity (ACN) of the principal surrogate approach. They assessed the conceptual setup of each approach in relation to the causal association framework of Joffe and Greene (2009). They quantify a causal association framework surrogate using the causal correlation of the potential treatment outcomes on S and T, they call this the individual causal association. They also investigated how the meta-analytical approach and ACN behaved under a range of scenarios with particular casual

relationships. They found that the ACN of principal stratification was very restrictive and suffered identifiability and transportability issues. Evaluation based on the meta-analytical approach generally is also satisfied under a causal assessment. However, positive causal assessments based on one trial (current practice) can fail causal meta-analytical assessments if heterogeneity between trials is high and causal effects are weak.

2.4.2 Causally classified approaches and the surrogate paradox

Vanderweele (2013) assessed the approaches classified in Joffe and Greene (2009) to see if they could identify paradoxical surrogates, see Table 2.1. Paradoxical surrogates occur when there is a positive treatment effect on the surrogate and a positive relationship between the surrogate and the true outcome but a negative treatment effect on the true outcome. As previously discussed (section 2.3.1.3), the occurrence of an unidentified paradoxical surrogate could have fatal consequences.

Vanderweele (2013) suggested that surrogates may be paradoxical in three scenarios; first, where there was a direct effect of the treatment on the true outcome which by definition circumvents the surrogate; second, there was a confounding influence; or finally, treatment may affect the true outcome and the surrogate but not for the same individuals, this they call “transitivity”.

Joffe and Greene (2009) evaluation approach	Can it identify the surrogate paradox?	
	Theoretically	Practically
PTE	No	No
Direct and indirect effects	Yes	No-identifiability issues
Principal stratification	Yes – not as part of the original proposition but can be under additional proposals Chen et al. (2007) and Ju and Geng (2010)	No-identifiability issues
Meta-analytical approach	Yes – not set up to do so - but can be done under proposals of Elliot et al. (2014)	Yes

Table 2.1: *Proposals of Joffe and Greene (2009) that can identify surrogate paradox*

The first of the approaches of Joffe and Greene (2009), the PTE, cannot recognise paradoxical surrogates. Theoretically, the direct and indirect effects approach can recognise them but identifiability issues mean this would be unfeasible in practice. Under conditions for "consistent" surrogates proposed by Chen et al. (2007) and Ju and Geng (2010) (see section 2.3.1.3) PS can theoretically identify the surrogate paradox but identifiability issues would again make this difficult in practice. The meta-analytical approach (and by extension the information theory approach) is not set up to identify the paradox however it could be used to do so.

Elliot et al. (2014) took these ideas forward and proposed measures based on the meta-analytical approach methodology which assess the risk of the surrogate paradox. These calculate the probability of the surrogate paradox occurring given beneficial treatment effects on the surrogate, and the size of beneficial treatment effect on the surrogate needed to prohibit the occurrence of the surrogate paradox.

2.5 Interdisciplinary approaches

The following approaches to surrogacy evaluation use methods from a range of areas of statistical methodology.

Endogeneity concerns a dependent variable, in our case the surrogate, being correlated with the error term through an unobserved confounder variable *U* (as discussed in section 2.4). The issue of the unobserved confounder, U, can be resolved through the use of an instrumental variable which is correlated with the dependent variable and dependent on *U* but not part of the explanatory model. Ghosh et al. (2010) proposed the investigation of the following structural equation model for a single trial:

$$T = \alpha_0 + \alpha S + \varepsilon_S \quad (18)$$

where α_0 and α are unknown intercept and treatment effect terms and ε_S is an error term with unknown distribution. Treatment, Z, is absent from (18) however in endogeneity it will contribute to the estimation of α if Z is an instrumental variable. To be instrumental it must fulfil three assumptions, one of which ties the concept of instrumentality to the evaluation of surrogacy and states that the effect of Z on T be fully explained by S (if Z is an instrumental variable it follows that S is a good surrogate for T.) They proposed a measure, related to the meta-analytical approach, and showed that it is causal under the instrumentality assumption on which it is strongly dependent. In the case where a mediation variable is controllable, Emsley et al. (2010) built two models and use endogeneity to aid identification. They assumed that a pre-randomisation covariate is instrumental to gain reliable estimates. However, this assumption has untestable components and relies on finding a suitable covariate.

Several authors have considered surrogate evaluation from a missing data point of view. Some highlighted the usefulness of surrogates to estimate parameters where the true outcome is incomplete (Chen et al., 2003b) and (Chen et al., 2008). Benda and Gerlinger (2007) investigated the use of sperm count as a surrogate for a binary pregnancy true outcome. Surrogate data was missing in the initial trial but recorded pregnancy rates for sperm count intervals were available. They found that the more information in the initial trial the greater precision they had in estimates of the level of pregnancy in the trial of interest. Korn et al. (2005a) and van Walraven et al. (2009) considered approaches where only trial level data is available, for multiple trials. The former focused on prediction of event rates between study groups which

are then used for comparison and the latter focused on trial level summaries that do not require individual level information. (Freedman, 2005) had some concerns about the approach in Korn et al. (2005a) which were further discussed in Korn et al. (2005b).

For multiple trials and binary outcomes, Baker et al. (2012) investigated surrogacy using a formal assessment of a surrogate's ability to predict treatment effect on the true outcome. They proposed a 95% prediction model incorporating an estimated random extrapolation error. Surrogates are assessed on the basis of a standard error multiplier; this compares the difference in the standard error calculated using the true outcome versus the standard error calculated using the prediction of true outcome based on the surrogate. A small standard error multiplier implies a good prediction model and therefore a useful surrogate.

2.6 Miscellaneous

Begg and Leung (2000) suggested there was a paradox in a Prentice based approach by Day and Duffy (1996). However, Baker et al. (2005) contested this assertion since contradictions exist in both authors calculations. MacKinnon et al. (2007) for mediation outcomes, proposed calculating $\beta - \gamma_Z$ from the single trial version of models (10) and (12). If this measure equals zero, the mediator fully explains the true outcome. Deslandes and Chevret (2007) proposed suitable joint models for two settings; first, repeated measures surrogates and time-to event true outcomes and second, multistate surrogates and a binary true outcomes. They then formally assessed surrogacy using a measure of PTE.

Event free survival is a commonly used surrogate which a composite made up of some interim outcome, for instance relapse, and the true outcome. Ghosh (2008a) argued that in this case "a practical definition of surrogacy is one in which a composite outcome based on both the surrogate [interim outcome] and the true outcome yields the same result as the true outcome". Based on this definition for time-to-event outcomes, the authors proposed an estimation of surrogacy using the correlation between treatment effects on the true outcome in relation to that on the composite outcome.

2.7 Surrogacy schemes

Schemas for assessing levels of surrogacy have been suggested. Qin et al. (2007) and Gilbert et al. (2008) proposed a 3 level schema for biological measures focusing on vaccine studies. With increasing validity, a 'correlate of risk' is a biological measure that correlates with the primary outcome, this is easily verified. A 'level one surrogate of protection' is a correlate of risk whose response to vaccine can be used to predict the vaccine response on the true outcome and a 'level two surrogate of protection' is a level one surrogate of protection which can be used for prediction in different populations or treatments. The authors conceded that level two surrogates of protection are extremely difficult to assess. The authors proposed a means of evaluating surrogates under this schema, for time-to-event outcomes, but experienced computational issues with both a principal surrogate and a meta-analytical approach. Dunning (2008) suggested that the framework should identify surrogates that 'quantitatively predict the efficacy of the vaccine' and proposed a means of doing so. Gilbert et al. (2009) argued that their approach already incorporated this.

Lassere et al. (2007b), Lassere et al. (2007a) and Lassere (2008) proposed a system of validating surrogates for setting up a clinical trial. Their scheme is based on biological aspects, quality of data, practical issues and statistical evaluation. Hence it incorporates statistical and non-statistical aspects of surrogacy which are equally important. The proposed surrogate is scored between one and fifteen. Proposed surrogates with low scores are classified as biomarkers (disease centred variables of biological or pathological process) those with high scores as patient outcomes (variables that reflect how a patient feels functions or survives).

2.8 General practical issues

Sarkar and Qu (2007) investigated the impact of measurement error in surrogacy evaluation using a measure of the PTE called the *excess relative odds* , where the surrogate is continuous and the true outcome is binary. They found that in the presence of measurement error the excess relative odds are biased. In the case of binary outcomes, Kassai et al. (2005) suggested the *diagnostic odds ratio* , a measure of sensitivity and specificity, to quantify measurement error but provide no detailed approach for its use.

As previously discussed correlation is a poor indication of a surrogate's worth. Baker and Kramer (2003) showed that, even where perfect correlation exists, a surrogate does not necessarily predict treatment effect on the true outcome. Wang et al. (2012) find that correlation can be influenced by many different factors and is therefore unsuitable for surrogate evaluation.

A supposed advantage of the use of surrogates is the possibility of smaller clinical trials. Baker and Kramer (2012) investigated the extrapolation error for binary outcomes, see section 2.5, and suggested that small studies based on a surrogate could suffer serious problems. Kramer (2013) proposed the measure *relative error* (RE_{MIX}) which estimates the validity of an *extrapolation assumption*. They found that, under Prentice and principal stratification approaches, the RE_{MIX} is dependent on the size of the sample. They stated that a smaller surrogate clinical trial is more likely to give misleading results than a larger one.

2.9 Discussion

The use of adequate surrogates benefits patients, clinicians, researchers and pharmaceutical companies alike. Therefore, researchers have worked towards a satisfactory approach to surrogacy evaluation. This systematic review was conducted to investigate and consider the available statistical approaches for evaluating surrogates.

I found that historic proposals, Prentice, PTE, relative effect and adjusted association provide useful and motivating conceptual ideas for evaluation. However, these have long since been deemed impractical and suffer serious bias due to adjusting for a surrogate which, as a post-randomisation variable, allows estimation of treatment effects on the true outcome to be influenced by confounders.

The meta-analytical approach uses multiple trials to more accurately predict treatment effect on the true outcome. This has been superseded by the information theory approach which in addition offers unified interpretation across different types of outcome. Alternative approaches, the principal stratification and direct and indirect effects, are also prominent in the literature. A number of other avenues have

also been investigated showing that development in this area is by no means stagnating.

I note that the main theoretical divergence in surrogacy evaluation approaches surrounds the need for causal validity. Principal stratification and direct and indirect effects approaches both provide this. Of the two, principal stratification has undergone the most development for applied use and has been suggested to be of more practical benefit (Joffe and Greene, 2009). I find that, though both these approaches are worthy and ambitious, they do not perform well upon practical application. Both take a counterfactuals approach and require assumptions to aid identifiability but cannot estimate effectively when the data do not follow the rigid framework imposed by these assumptions. Researchers may face various quandaries over balancing levels of bias, due to the use of invalid assumptions, and quality of estimation due to inability to identify parameters. In summary, principal stratification and direct and indirect effects are valuable and stimulating approaches to surrogate outcome evaluation, chiefly because they examine causal effects. However, they are both somewhat in their infancy in regard to practical developments and, thus far, no consistent and established approaches have gained prominence.

I find, on balance, that the information theory approach is currently the preferred approach for assessing surrogacy. In comparison to its predecessor, the meta-analytical approach, it provides a unified interpretation across settings, does not rely on interpretations at the latent level, and computational issues regarding the use of joint models are avoided. Both the meta-analytical and information theory approaches avoid the need to fully assess causal pathways at the trial level while still retaining a causal interpretation. This is because it assesses the relationship between treatment effect on the true outcome and treatment effect on the surrogate, rather than evaluating whether treatment acts on the true endpoint by acting on the surrogate. The latter approach provides a more thorough causal investigation but is less worthwhile overall as it leads to serious practical issues regarding multi-faceted treatment mechanisms and outside influences (Joffe and Greene, 2009). Despite these advantages, information theory provides no causal interpretation at the individual level. However, causality is not the main focus of information theory,

which assesses the value of the surrogate as a source of information. The multi trial aspect of both approaches provides the key advantage in that it "deals, albeit imperfectly, with transportability" (Joffe, 2011). In other words, if the same effect is seen across trials this provides evidence that this effect will also be present in a new trial assuming it is performed in the same population. This means it can inform on the ability of the surrogate to predict treatment effects on the true outcome which is a key aim of surrogacy evaluation and is not currently a central feature of the causal approaches (Pearl, 2011).

Even though the information theory approach offers no causal interpretation at the individual level this negative is outweighed by its provision of a sophisticated approach to surrogacy evaluation that is straightforward to apply in practice, provides researchers with the pertinent information and has limited computational or interpretational issues.

Finally, impressive as the surrogacy evaluation literature is, Prentice (1989) noted that surrogacy evaluation applies at most to treatments that act through the same mechanism of action. None of the methods presented in this chapter has proposed a solution to the evaluation of surrogates for different classes of treatments.

This chapter has provided a comprehensive review of the statistical developments and current methodologies in the field of surrogate evaluation. The vast array and scope of the described publications demonstrate a continuing strong interest in developing valid surrogate evaluation approaches.

Chapter 3. Extension of information theory for a binary surrogate and ordinal true outcome: Methodology

As highlighted in the systematic review (Chapter 2), the most appealing approach to surrogacy evaluation from both a practical and theoretical perspective is the information theory approach. Practically speaking: it requires no stringent assumptions; suffers no serious computational difficulties; and provides a consistent approach across settings. From a theoretical perspective, it provides the pertinent information required and a causal interpretation at the trial level. Furthermore, it provides information on how well a surrogate is likely to perform as a predictor of treatment effect on the true outcome in a new trial. This is a fundamental aim of surrogacy and one which is not a principal feature of other well established approaches.

Given the strong arguments in favour of the information theory approach I extended this methodology to the case where either the surrogate, the true outcome, or both are ordinal. Previous authors had suggested the means of doing this for the meta-analytical approach (Burzykowski et al., 2003) but none are currently available for the information theory approach. Hence, researchers in clinical areas where ordinal outcomes are used will be able to evaluate surrogacy.

In this chapter I: outline the information theory approach in detail; show how this has been extended to the case of a binary surrogate and ordinal true outcome; present confidence intervals for information theory based measures; and outline my approach to avoiding bias where separation occurs.

The notation of the following sections is as follows. Y and X represent two random variables, either discrete or continuous depending on the context. In the discrete context, Y and X have values $k_b,\ b \in (1, \dots, m_y)$ and $k_d, d \in (1, \dots, m_x)$ and probabilities of occurrence of each value p_b and p_d respectively. I represent a putative surrogate as S, treatment as Z and the true outcome as T. In the multi trial context there

are i=1,2,…,N trials, and j=1,2,….,n_i patients per trial. $N_T = \sum_i n_i$, the total number of patients in all trials. W is the number of categories in the ordinal true outcome.

3.1 The meta-analytical approach: general truths

The interpretation of individual level surrogacy under the meta-analytical approach was different across different types of outcome. The information theory approach was developed to resolve these inconsistency issues. The meta-analytical and information theory approaches are very closely linked and analogous in many respects. Therefore, a lot of the work in publications for the meta-analytical approach are relevant to that under the information theory approach.

3.2 The information theory approach

Information theory uses the central concept of entropy to measure the "information, choice and uncertainty" in a random variable (Shannon and Weaver, 1948). Entropy measures how uncertain one is about the outcome of individual draws from a Markov process. Taking the example of a coin toss, where a coin is heavily unfair towards one particular outcome, each new flip of the coin does not provide much information. Since one outcome is a great deal more likely than the other we already have a strong idea what the outcome will be. On the other hand, a balanced coin is much less easy to predict and therefore each outcome of the coin toss is providing more information and hence has higher entropy.

Mathematically speaking, in the discrete case, entropy can be represented as $H(Y) = -\sum_b p_b \log(p_b)$, where Y is a discrete random variable with values $k_1, k_2, \ldots., k_{m_y}$ and probabilities $p_1, p_2, \ldots., p_{m_y}$ respectively (Shannon and Weaver, 1948). Entropy has a number of useful properties. For example, H=0 if and only if all but one probability takes the value zero (in other words, it has only one possible outcome), otherwise H is always positive. This can be demonstrated using the coin toss example: if the probability of a tail is one and a head is zero the new value of a toss will provide no information as we are 100% certain of the result. If a head has any probability larger than zero, a new toss will provide information and entropy will be larger than zero.

Through consideration of the case of the joint entropy of two random variables Y and X, where $p_{b,d}$ is the joint probability of Y=b and X=d, $H(Y,X) = -\sum_{b,d} p_{b,d} \log(p_{b,d})$ and their conditional entropy $H(Y|X) = -\sum_{bd} p_{b|d} \log(p_{b|d})$ it can be shown that $H(Y) \geq H(Y|X)$. This means that the entropy of Y is bigger or equal to the entropy of Y given X. Some intuitive conclusions follow from this: $H(Y) = H(Y|X)$ if and only if X and Y are independent, see Shannon and Weaver (1948); the uncertainty in Y is never increased by knowledge of X; and H(Y) is invariant by bijective transformations of Y. A full list and proofs of these and other properties of entropy can be found in Shannon and Weaver (1948).

A concept of fundamental importance is called the mutual information. This is defined as $I(X,Y) = H(Y) - H(Y|X)$ and is interpreted as the amount of uncertainty in Y removed if X is known. These concepts are useful in surrogate evaluation as we are interested in the amount of information on T removed through the knowledge of S at the individual level. At the trial level we are interested in the amount of information of the treatment effects on T removed through the knowledge of treatment effects on S.

Alonso and Molenberghs (2007) used these concepts and presented an information theory approach to surrogacy evaluation. Their approach was based on information theory concepts for continuous outcomes. Differential entropy was used to measure the entropy of continuous random variables. Assume Y is now a continuous random variable and f_y is the corresponding density function, we define differential entropy as $h_d(Y) = -\int f_y(y) \log\left(f_y(y)\right) dy$. This holds most but not all of the properties of the discrete case. Key differences are that $h_d(Y)$ does change through transformations of Y and it can take negative values. Again, considering the case of two continuous variables X and Y and with joint density f_{yx}, the conditional differential entropy, $h_d(Y|X)$, and the mutual information, $I(X,Y) = h_d(Y) - h_d(Y|X)$, are defined in an analogous manner to the discrete case.

The information theory measure of surrogacy was based on another more useful concept in the continuous setting, the entropy power, obtained by maximising the

entropy of a continuous random variable, defined as $EP(Y) = \frac{1}{(2\pi e)} e^{2h(Y)}$. The differential entropy of a normally distributed continuous random variable is $H(Y) = \frac{1}{2}\log(2\pi e\sigma)$ meaning that $EP(Y) = \sigma^2$. This suggests that for normally distributed variables information and variability are equivalent. However, in practice *EP(Y)* is larger than *Var(Y)* if the continuous variable is not normally distributed.

3.3 Individual level surrogacy: information theory

At the individual level Alonso and Molenberghs (2007) proposed an information theory surrogate evaluation measure:

$$R_h^2 = \frac{EP(T) - EP(T|S)}{EP(T)} \qquad 3.1$$

Where, $EP(T)$ is the entropy power of T and $EP(T|S)$ is the entropy power of T given S. This can be interpreted as the amount of uncertainty in the true outcome T removed when S is known. R_h^2 has useful properties including that: it is linked to the mutual information through $R_h^2 = 1 - e^{2I(S,T)}$; R_h^2 is invariant by bijective transformations of S and T; for continuous variables there exists a deterministic relationship between S and T. Finally, $R_h^2 = 0$ if and only if T and S are independent, this result would suggest a poor surrogate that explains none of the uncertainty in T as you would expect if S and T were independent.

Alonso and Molenberghs (2007) were keen to align the information theory approach to the meta-analytical approach by imposing a multi-trial framework. This was so that the information theory approach could tackle the issue of transportability and be used to predict the treatment effect on T in a new trial. However, in the multi-trial framework there were an uncountable number of options for the choice of summary parameter to be used to calculate surrogacy.

In equation 3.2, if we have N trials we also have N_q possible values of $R_{h_i}^2$, the R_h^2 for the i^{th} trial, since we can cluster trials depending on q different characteristics. Hence, there are many different choices for the parameter ϑ_i and an uncountable number of summary measures we could use to calculate R_h^2 in the multi-trial setting.

$$R_h^2 = \sum_{i=1}^{N_q} \vartheta_i R_{h_i}^2 = 1 - \sum_{i=1}^{N_q} \vartheta_i e^{-2I_i(s_i,t_i)}, \qquad 3.2$$

$$\vartheta_i > 0 \; \forall \; i, \; \sum_{i=1}^{N_q} \vartheta_i = 1$$

Alonso and Molenberghs (2007) proposed a number of options but highlight the Likelihood Reduction Factor (LRF) as a good candidate to calculate R_h^2 in the multi-trial setting. The LRF is particularly useful as it ranges in the unit interval and has a consistent interpretation across outcome types, for example the continuous, binary and time-to-event settings. As noted in section 2.2.5, the reason that the information theory approach was created was that its predecessor, the meta-analytical approach, does not provide a consistent interpretation at the individual level.

In the following sections I introduce the LRF for continuous outcomes at the individual level and then show how this can be extended to the case of a binary surrogate and ordinal true outcome.

3.3.1 Individual level: likelihood reduction factor

The LRF was introduced as a means of evaluating surrogacy by Alonso et. al. (2006). Here I present the fixed effects LRF at the individual level, for the continuous-continuous setting. The LRF is based on the amount of information gained about the true outcome after accounting for the surrogate. These authors proposed modelling 3.3 and 3.4 for each trial *i*, these models regress the true outcome on treatment with and without adjustment for the surrogate respectively:

$$T_{ij} = \mu_i + \beta_i Z_{ij} + \varepsilon_{T_{ij}} \qquad 3.3$$

$$T_{ij} = \theta_{0_i} + \theta_{1_i} Z_{ij} + \theta_{2_i} S_{ij} + \varepsilon_{ij} \qquad 3.4$$

where: θ_{0_i} and μ_i are intercept parameters with and without adjustment for the surrogate; β_i is the treatment effect parameter for the true outcome; θ_{1_i} and θ_{2_i} are treatment and surrogate parameters for the model with adjustment for the surrogate. The difference in the amount of information on the true outcome gained from the surrogate is calculated via the difference in the log-likelihood between 3.3 and 3.4

which is formally expressed as G_i^2, for each trial i. LL_0 is always the log-likelihood for the unsaturated model, in this case 3.3, and LL_1 for the saturated model, 3.4, for trial i. $G_i^2 = 2 * (LL_1 - LL_0)$. The LRF is then calculated:

$$LRF = R_h^2 = 1 - \frac{1}{N}\sum_i \exp\left(-\frac{G_i^2}{n_i}\right) \quad 3.5$$

The authors stated that the LRF reduces to the R^2_{indiv} measure of the meta-analytical approach (see section 2.2.1) where outcomes are normally distributed. We will use the LRF to calculate the R_h^2 at the individual level for the binary-ordinal setting.

3.3.2 Individual level: binary-ordinal

At the individual level I applied the LRF in the binary-ordinal setting in the same manner as in the continuous case using (13) however this was based on the difference in $G^2 = 2 * (LL_1 - LL_0)$ of the following proportional odds models:

$$logit\big(P\big[T_{ij} \leq w\big]\big) = \mu_{T_{w_i}} + \beta_i Z_{ij} \quad 3.6$$

$$logit(P[T_{ij} \leq w]) = \theta_{0_{w_i}} + \theta_{1_i} Z_{ij} + \theta_{2_i} S_{ij} \quad 3.7$$

Where, $w = 1, \ldots, W - 1$, and W is the number of categories in the ordinal true outcome. $\mu_{T_{w_i}}$ and $\theta_{0_{w_i}}$ are intercept parameters for each cut point of the ordinal true outcome for each trial, β_i and θ_{1_i} are treatment and θ_{2_i} surrogate parameters.

Again, the LRF is based on the difference in the amount of information gained on the true outcome with and without the surrogate for each trial.

However, in the case of discrete outcomes and a family of conditional models the LRF is bounded above by a number strictly less than one. Consider the case of the calculation of information gain, G^2, between an intercept only model and a saturated model with one additional variable. With log-likelihoods respectively of LL_0 and LL_1 and $G^2 = 2(LL_1 - LL_0)$. In the case of continuous responses, the probability of a specific value in the random variable is zero. However, in the discrete case all values of the response variable will have a non-negative probability hence Kent (1983) stated

that $LL_0 \leq LL_1 \leq 0$. Therefore, Kent (1983) further stated that for discrete response variables $G^2 \leq -LL_0$. Meaning that G^2 is less than or equal to the total amount of information in the response variable, as determined by the log-likelihood intercept only model of that variable.

In the case of surrogacy evaluation in principle this response variable would be the true outcome T. And ideally, $G^2 \leq -LL_T$, where LL_T is the log-likelihood of the intercept only model, $(logit(P[T_{ij} \leq w]) = \theta_3)$ and LL_T represents the total information present in T. $G^2 \leq -LL_T$ would be a desirable result since it would mean that it would be mathematically possible for the surrogate to explain all the information that is present in the true outcome. And, therefore, we could determine the amount of information in T that the surrogate explains. However, in the case of individual level surrogacy because both models are conditional on treatment it is not possible for the surrogate to explain all the information in T because $G^2 \leq -LL_T$ does not hold.

Consider the two nested conditional models of the individual level, 3.6 and 3.7. These are nested conditional models as both are conditional on treatment, and 3.7 is the saturated model as it is also conditional on S. Let's define the log-likelihoods of these conditional models as $LL_{T|Z}$ and $LL_{T|Z,S}$ respectively. In this case G^2 is bounded above by the information in the true outcome given the shared conditional variable, Z, rather than by the information in the true outcome alone. Mathematically this is $G^2 \leq -LL_{T|Z}$ rather than $G^2 \leq -LL_T$, see Kent (1983). Therefore, at the individual level, G^2 is bounded above by a value which is less than the full information in the true outcome. That is unless the conditional variable, Z, explains none of the variability. Since R_h^2 is based on the G^2 this in turn means that R_h^2 is bounded above by a number less than 1, $R_h^2 \leq 1 - e^{2LL_{T|S}}$, as shown in Alonso and Molenberghs (2007).

Hence, Alonso and Molenberghs (2007) proposed rescaling R_h^2 so that is composed in reference to the information in T alone rather than T|Z and is bounded above by 1:

$$\widehat{R_h^2} = \frac{R_h^2}{1 - e^{2LL_T}} \qquad 3.8$$

They do this by rescaling R_h^2 using the information in T alone as calculated by LL_T, see 1.8. Despite the rescaling the LRF has a consistent interpretation in this setting as well as that described previously.

3.3.3 Individual level: modelling methods binary-ordinal

For R_h^2 the LRF between the model with and without adjustment for the surrogate for each trial, see 3.3.2. However, an alternative procedure is to calculate only two models in total incorporating all trials. This can be done by using trial as a fixed effect variable in the model, as opposed to the two models 3.6 and 3.7 for each trial. Here:

$$logit\left(P\left[T_{ij} \leq w\right]\right) = \mu_{T_w}^0 + \mu_{T_i} trial_{ij} + \beta_i (trial_{ij} * Z_{ij}) \qquad 3.9$$

$$logit(P[T_{ij} \leq w]) = \qquad 3.10$$

$$\theta_{0_w}^0 + \theta_{0_i} trial_{ij} + \theta_{1_i} (trial_{ij} * Z_{ij}) + \theta_{2_i} (trial_{ij} * S_{ij})$$

Where: $\mu_{T_w}^0$ and $\theta_{0_w}^0$ are fixed intercepts for each of the *W-1* cut points of the ordinal true outcome; μ_{T_i} and θ_{0_i} are the trial specific shifts of the set of intercepts; β_i and θ_{1_i} are treatment and θ_{2_i} surrogate parameters for each trial *i*.

This method would lead to an LRF based on the G^2 of only 3.9 and 3.10. Rather than the G_i^2 of 3.6 against 3.7 for each trial i which are then summed over all trials in the approach outlined in section 3.3.2. This two model only method was used to calculate individual level surrogacy in previous information theory publications in the binary-binary setting (Tilahun et al., 2008) and binary-continuous setting (Pryseley et al., 2007); according to their freely available software (I-Biostat, 2015). In this case, the LRF would be composed as in 3.11 as opposed to (13):

$$LRF = 1 - exp\left(-\frac{G^2}{N_T}\right) \qquad 3.11$$

The LRF in equation 3.11 for the individual level is only valid under the assumption that the "association between both variables is constant over trials" (Alonso et al., 2004). This assumption may not hold in real life situations. Furthermore, practically

speaking, using this method in the ordinal setting, would lead to serious issues of over-fitting because of the interaction term $trial_{ij} * S_{ij}$ in the second model (equation 3.10). Conversely, basing the LRF on (13) using separate models for each trial *i* is more in keeping with the multi-trial philosophy of the information theory approach; modelling issues are limited; and potentially invalid assumptions are not necessary. Hence, we will use separate models for each trial and apply the LRF using (13).

3.4 Trial level surrogacy: information theory

At the trial level we are interested in the treatment effects on the surrogate in relation to the treatment effects on the true outcome. In order to calculate this Alonso and Molenberghs (2007) proposed a two stage approach. At the first stage we obtain the treatment effects for each trial on the surrogate and true outcome, α_i and β_i respectively. This is done by regressing the surrogate and true outcome on treatment in the models 3.12 and 3.13. First I present the models required for the continuous-continuous case:

$$S_{ij} = \mu_{S_i} + \alpha_i Z_{ij} + \varepsilon_{ij} \qquad 3.12$$

$$T_{ij} = \mu_{T_i} + \beta_i Z_{ij} + \varepsilon_{ij} \qquad 3.13$$

Where μ_{S_i} and μ_{T_i} are intercept parameters and α_i and β_i are treatment effect estimates for S and T respectively for each trial i. Using the treatment effect estimates we calculate the information theory surrogacy measure R^2_{ht} through 3.14.

$$R^2_{ht} = \frac{EP(\beta_i) - EP(\beta_i|\alpha_i)}{EP(\beta_i)} \qquad 3.14$$

Where $EP(\beta_i)$ is the entropy power of treatment effects on the true outcome and $EP(\beta_i|\alpha_i)$ is the entropy power of treatment effects on the true outcome given those on the surrogate. R^2_{ht} can be interpreted as the amount of uncertainty in the treatment effect on T removed through knowledge of the treatment effect on S. If $R^2_{ht} \approx 1$ then the treatment effects on the surrogate explains a high level of the uncertainty in the treatment effects on the true outcome.

In the following sections I introduce the LRF for continuous outcomes at the trial level and then show how this can be extended to the case of a binary surrogate and ordinal true outcome.

3.4.1 Trial level: likelihood reduction factor

The LRF can be applied to calculate R_{ht}^2 in the continuous-continuous case. In order to do this we model 3.12 and 3.13, the surrogate and true outcome regressed on the treatment respectively. These models provide treatment effect estimates for both S and T for each trial, these are represented by the parameters α_i and β_i respectively. At the second stage two further models are required: the null model (3.15) of treatment effects on the true outcome for each trial; and the treatment effects on the true outcome regressed on the intercepts and treatment effects of the surrogate for each trial (3.16).

$$\widehat{\beta_i} = \gamma_3 + \varepsilon_i \qquad 3.15$$

$$\widehat{\beta_i} = \gamma_0 + \gamma_1 \widehat{\mu_{S_i}} + \gamma_2 \widehat{\alpha_i} + \varepsilon_i, \qquad 3.16$$

where, γ_3 and γ_0 are the intercept parameters with and without adjustment for the surrogate, γ_1 and γ_2 are the parameters for the surrogate intercept and treatment effect estimates provided from stage one. The difference in the -2*log-likelihood between these two models can then be calculated and the LRF applied as in 3.17.

$$LRF = \widehat{R_{ht}^2} = 1 - \exp\left(-\frac{G^2}{N_T}\right) \qquad 3.17$$

3.4.2 Trial level: binary-ordinal

In the binary-ordinal setting the only difference in the approach is in the models used at the first stage. Here a generalised linear model is fitted for the binary surrogate regressed on treatment, in 3.18. A proportional odds model is fitted for the ordinal true outcome regressed on treatment, in 3.19. This is done to estimate the treatment effects β_i and α_i.

$$logit[P(S_{ij} = 1)] = \mu_{S_i} + \alpha_i Z_{ij} \quad 3.18$$

$$logit[P(T_{ij} \leq w)] = \mu_{T_{w_i}} + \beta_i Z_{ij} \quad 3.19$$

Where, $w = 1, \ldots, W-1$, and W is the number of categories in the ordinal true outcome, $\mu_{T_{w_i}}$ is the set of intercept parameters for each *W-1* cut point of the ordinal true outcome and all other parameters are the same as the continuous case. The second stage models 3.15 and 3.16 can be fitted in the same manner as in the continuous setting using the parameters of 3.18 and 3.19, and the LRF applied as in 3.17. The LRF gives consistent results at the trial level for both the continuous setting and the binary-ordinal setting.

3.4.3 Trial level: modelling methods

The calculation of 3.12 and 3.13 can be done in one of two ways in order to gain estimates of β_i and α_i. First, as described above, by modelling the surrogate/true outcome regressed on the treatment separately for each trial, see equations 3.18 and 3.19. Or secondly, by using two models only, regressing surrogate or true outcome on the terms $trial_{ij}$ and the interaction $trial_{ij} * Z_{ij}$, in a similar manner to that discussed for individual level surrogacy (see section 3.3.3). Either method will return the required parameter estimates β_i, μ_{S_i} and α_i for each trial and since the latter method is less laborious this will be used hereafter.

3.4.4 Trial level: discussion

3.4.4.1 Weighting by trial size

Given that the R^2_{ht} is calculated in two stages the LRF is based only on one G^2. This leads to an issue at the trial level since the two stage approach does not adequately take into account the differences between trials. The only way to achieve this is to use the full bivariate meta-analytical approach. However, as discussed in section 2.2.2 this is extremely computationally burdensome even in the continuous-continuous setting. (The computational issue was the reason for the development of the two stage approach.) As a partial remedy to this problem in the two stage approach, in the context of the meta-analytical approach Tibaldi et al. (2003a) suggested adjusting the analysis

to account for trial size. This could be an equally valid technique in the information theory setting.

It is important to take account of trial size in the analysis as smaller trials will have less ability to estimate treatment effects accurately. These treatment effect estimates are used at the second stage of the analysis, see 3.15 and 3.16. If the models at the second stage of modelling are unweighted then these smaller trial's treatment estimates will contribute just as much to the model as larger trial's estimates, which are likely to be more precise. Weighting by trial size should enable the model to put more emphasis on estimates that are more reliable.

In order to achieve this at the second stage of modelling a weighting term can be added to the linear models, 3.15 and 3.16, based on the exact trial size, therefore trials that are larger contribute more to the model. Weighting can be applied to linear models in R using the weight term in the function `lm`. The `weight` term tells the model to conduct weighted least squares which in this case minimises $\sum_i weight_i * e^2$ where $weight_i$ is the set weights for each trial *i, R core team (2016).* If $weight_i$ in this case is set to the exact size of the trial *i* then each trials contribution at the second stage is weighted according to the size of the trial with larger trials contributing more to the analysis.

3.4.4.2 Connections to meta-analytical approach

It is interesting to note that the information theory trial level surrogacy is based on the same model at the second stage, equation 3.16, as that used under the simplification of Tibaldi et al. (2003b) at the second stage of the meta-analytical approach (see section 2.2.1). Tibaldi et al. (2003b) uses the coefficient of determination of this model as the measure of surrogacy rather than information gain but the two approaches should give identical results in practice. This is an example of the close links between the information theory and meta-analytical approaches. It also demonstrates why it can be reasonable to compare the theory and simulation results from publications under the meta-analytical approach to the ones from information theory.

3.5 Confidence intervals: binary-ordinal

Confidence intervals are calculated based on G^2, of the LRF see section 3.3.1, and therefore on the log-likelihoods of two models. Though the models might differ between settings, i.e. generalized linear or proportional odds models, the form of the variable utilised G^2 is the same. Therefore, there are no differences at all between settings in how the confidence intervals are calculated or in their assumptions. The one exception is in the rescaling of intervals for discrete outcomes at the individual level, which will be discussed in 3.5.2, however this has no impact on the interpretation of the confidence intervals or their assumptions.

Therefore, confidence intervals give consistent interpretations between settings in much the same way as the LRF, which is advantageous as consistency was the main driving force for the development of the information theory approach.

The LRF of Alonso and Molenberghs (2007) was based on the ideas of Kent (1983) who utilised the non-central χ^2 distribution in order to provide confidence intervals for the information gain G^2. Kent (1983) proposed different intervals for small and large information gain. A distinction between large and small was not provided but apparently the asymptotics for the small confidence intervals are 'more useful', therefore, I used these intervals.

The central χ^2 distribution is usually referred to simply as the χ^2 distribution. We can think of the central χ^2 distribution as based on the null hypothesis. In this case the null hypothesis, H0: $G^2 = 0$, is that the two models used to calculate the LRF have the same log-likelihood, with the alternative hypothesis H1: $G^2 \neq 0$. G^2 is the gain in information calculated using the -2*log-likelihood ratio test between two models. The non-central χ^2 distribution represents all the possible distributions under the alternative hypothesis, H1: $G^2 \neq 0$, where there are deviations from the null, $G^2 = 0$. To be clear, a hypothesis test at this point is not being administered, the non-central χ^2 distribution can just be conceptualised under this terminology. Two important components of the confidence intervals proposed by Kent (1983) are defined in the following:

$$P\left[\chi_p^2(\gamma_{p:\alpha}(G^2)) \geq G^2\right] = \alpha \text{ and } P\left[\chi_p^2(\delta_{p:\alpha}(G^2)) \leq G^2\right] = \alpha \qquad 3.20$$

Where $\gamma_{p:\alpha}$ and $\delta_{p:\alpha}$ are draws from the non-central χ^2 distribution where an information gain of G^2 is observed. These parameters are based on the significance level α. Here p represents the degrees of freedom, this is equal to the number of additional parameters in the fuller model. At the individual level, model 3.7 incorporates the binary surrogate in addition to the parameters of 3.6 which is one additional degree of freedom, hence p=1. The equalities in 3.20 return the value of the χ^2 distribution where the probability of observing values at or larger/lower than G^2 equal α. In other words, in the case of a 95% confidence interval, these are the 2.5th and 97.5th percentiles of the non-central χ^2 distribution that deviates from the null hypothesis of no difference by G^2. This is unless:

$$P\left[\chi_p^2(0) \geq G^2\right] > \alpha \text{ then } \gamma_{p:\alpha}(G^2) = 0 \qquad 3.21$$

$\chi_p^2(0)$ is the central χ^2 distribution or χ^2 under the null hypothesis, H0: $G^2 = 0$. We wish to determine the probability of G^2or one more extreme occurring under the null hypotheses. In other words conduct a significance test under the null hypothesis of G^2=0. If the p-value is larger than α we cannot reject the null hypothesis that there is no information gain and the lower limit of the confidence interval is set to equal zero. This makes sense intuitively because if the confidence interval contains zero we cannot reject the null hypothesis that there is no difference.

Kent (1983) suggest a (1-alpha)% confidence interval for an estimate of G^2calculated using the difference in 2*log-likelihood between the two models and represented as $\widehat{G^2}$:

$$\left[\frac{\mu\gamma_{1:\alpha/2}\left(\widehat{G^2}/\mu\right)}{n_k}, \frac{\mu\delta_{1:\alpha/2}\left(\widehat{G^2}/\mu\right)}{n_k}\right] \qquad 3.22$$

Here n_k is the number of observations the models are based on. Suppose we have $\hat{\theta}$ which is the maximum likelihood estimate under the parameter space $\Theta_1$of the larger model. The log-likelihood ratio test is applied. Then μ is equal to one if 'the true density function belongs to $\{f(x; y; \theta|\theta\epsilon\Theta_1\}$' Bergman and Holmquist (2012). This

means that we assume that one combination of variables constitutes the correct model in equation 3.7 in the individual level case or 3.16 at the trial level. Since we must compose our models the way described in order to investigate surrogacy this is a valid assumption in this setting. Therefore, our confidence intervals became:

$$\left[\frac{\gamma_{1:\alpha/2}\left(\widehat{G^2}\right)}{n_k}, \frac{\delta_{1:\alpha/2}\left(\widehat{G^2}\right)}{n_k}\right] \quad 3.23$$

where $\gamma_{1:\alpha/2}\left(\widehat{G^2}\right)$ and $\delta_{1:\alpha/2}\left(\widehat{G^2}\right)$ were calculated as described above. In order to provide confidence intervals for the LRF as opposed to the G^2, Kent (1983) proposed that they should be converted to:

$$\left[1 - exp\left(-\frac{\gamma_{1:\alpha/2}\left(\widehat{G^2}\right)}{n_k}\right), 1 - \exp\left(-\frac{\delta_{1:\alpha/2}\left(\widehat{G^2}\right)}{n_k}\right)\right] \quad 3.24$$

3.5.1 Confidence intervals: trial level

At the trial level the confidence intervals could be applied by calculating the bounds of the interval according to 3.24, where $n_k = N_T$ the total number of patients for all trials.

3.5.2 Confidence intervals: individual level

Confidence intervals for R_h^2 at the individual level have multiple ${G_i}^2$ values, for each trial i, therefore converting the intervals to those for the LRF was more complicated. I decided to follow the setup of the LRF of Alonso et al. (2006) for R_h^2 to calculate these:

$$\left[1 - \frac{1}{N}\sum_i exp\left(-\frac{\gamma^i_{1:\alpha/2}\left(\widehat{G^2}\right)}{n_i}\right), 1 - \frac{1}{N}\sum_i exp\left(-\frac{\delta^i_{1:\alpha/2}\left(\widehat{G^2}\right)}{n_i}\right)\right] \quad 3.25$$

The confidence intervals at the individual level also require adjustment in the same manner as R_h^2 since they are bounded above by a number less than one, see section 3.3.2.

3.6 Separation: binary-ordinal

In the case of categorical variables, separation and quasi-complete separation can occur: in these instances, maximum likelihood estimates are not unique. In this section I discuss: how complete and quasi-complete separation occurs both in the binary and ordinal case; how this affects the calculation of R^2_{ht}; and discuss a solution to this issue.

3.6.1 Separation: binary

Consider the case of two binary variables, as in the calculation of R^2_{ht} at the first stage of modelling where a binary surrogate is regressed on a binary treatment variable for a certain trial (see 3.18). Complete and quasi-complete separation is related to the occurrence of zero cells in a cross-tabulation of these two variables.

The occurrence of no separation is shown in Table 3.1.

		Treatment	**Placebo**
Surrogate	**Y**	A≠0	B≠0
	N	C≠0	D≠0

Table 3.1: No separation when comparing binary outcomes, no zero cells.

Complete separation occurs when a binary variable X can perfectly predict Y as represented in Table 3.2:

		Treatment	Placebo
Surrogate	**Y**	A≠0	0
	N	0	D≠0

		Treatment	Placebo
Surrogate	**Y**	0	B≠0
	N	C≠0	0

Table 3.2 *Complete separation of two binary variables*

In comparison the more common issue of quasi-complete separation occurs if any of the cells includes a zero value. For example, the following case:

		Treatment	Placebo
Surrogate	**Y**	A≠0	B≠0
	N	C≠0	0

Table 3.3: *An example of quasi-complete separation of two binary variables*

In both complete and quasi-complete separation the maximum likelihood is bounded above by a number less than zero however it has no maximum (Allison, 2008). A graphical representation of this can be seen in Figure 3, which is an increasing function but has no maximum.

There is no maximum because of the way in which the maximum likelihood of the parameter of interest is calculated. In the case of two binary variables this is as follows:

$$\hat{\varphi} = \ln\left(\frac{A * D}{B * C}\right) \qquad 3.26$$

Here we can see that if any or more than one of A, B, C or D equal zero issues occur. If a zero occurs on the denominator then $\hat{\varphi}=\infty$ which is undefined, if a zero occurs on the numerator then $\hat{\varphi}=\ln(0)$ which is also undefined (Allison, 2008).

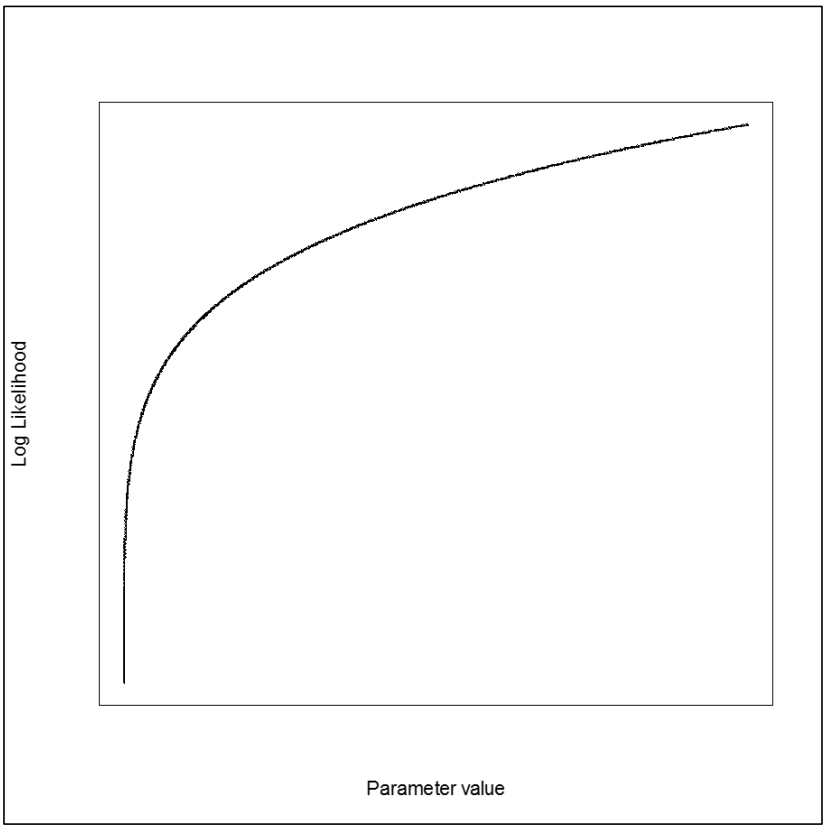


Figure 3*: Hypothetical example of the log-likelihood of a parameter which suffers separation*

3.6.2 Separation: ordinal

At the first stage of calculating R_{ht}^2 in the binary-ordinal setting an ordinal true outcome is regressed on a binary treatment variable for a certain trial (see 3.18). An ordinal outcome against a binary treatment indicator can also suffer separation. This has the same consequences on the calculation of R_{ht}^2 as in the binary case. The description of separation for ordinal variables I present will refer back to the case of two binary variables. In the following tables any character in a cell of a table represents a number greater than zero and hence is not a zero cell.

Here I present the definition of separation for ordinal outcomes as given by Agresti (2014). However, he labelled all possible occurrences of separation in the ordinal case as quasi-complete separation.

Imagine you collapse the categories of an ordinal variable into a binary variable at each possible threshold. For each collapse, if one or more of the cells in the two by two crosstab is zero then quasi-complete separation exists (Agresti, 2014).

In the example in Table 3.4, if you dichotomise the seven point scale into binary groups at any threshold of the scale the resultant crosstabs contain zero cells, and would look like Table 3.3. This is an example of quasi-complete separation since at each ordinal threshold separation occurs.

1	**2**	**3**	**4**	**5**	**6**	**7**
$\boldsymbol{A_1}$	B_1	C_1	D_1	E_1	F_1	G_1
$\boldsymbol{A_2}$	B_2=0	0	0	0	0	0

Table 3.4 quasi-complete separation example for ordinal variable

However if $B_2 \neq 0$ and you were to dichotomise at one you get Table 3.1. In this case, since dichotomisation at at least one cut point gives no separation in the binary case there is no quasi-complete separation.

Another way in which the ordinal variable may have quasi-complete separation is if there "exists a pair of rows for which all observations on one row never fall above any observation in the other row" (Agresti, 2014).

In Table 3.5 we see an example of this situation where none of the values of the categories recorded in row one are higher than any recorded in in row two, even though some values are the same. In other words, there is no overlap of at least two categories between rows. This type of separation equates closely to complete separation in the binary case and will be referred to as such for ordinal outcomes in the remainder of this thesis.

1	**2**	**3**	**4**	**5**	**6**	**7**
$\boldsymbol{A_1}$	B_1	C_1	D_1	0	0	0
0	0	0	D_2	E_2	F_2	G_2

Table 3.5: quasi-complete separation example for ordinal variable

3.6.3 Impact of separation on surrogate evaluation

Now I will discuss how separation affects surrogacy evaluation. I start by describing the way that functions for proportional odds or generalised linear models behave in R, SAS and other statistical programs when separation occurs.

According to Allison (2008) the typical scenario is that the model will try to converge but the occurrence of zeros will prevent this since there is no unique-maximum likelihood of the affected parameters. The model will go through several iterations attempting to converge. Upon each iteration the affected parameter estimate will increase and this will continue until a fixed iteration limit is exceeded. At this point the affected parameter estimate will typically be very large and its standard error may be extremely large. Statistical software generally tends to be poor at reporting this issue through error messages.

In the case of surrogate evaluation at the trial level we use a two stage approach. At the first stage two binary (S|Z) or an ordinal and a binary variable (T|Z) are regressed on one another (equations 3.18 and 3.19) for each trial. This is with the aim of returning treatment effect estimates on the binary surrogate and ordinal true outcome, however if separation occurs for a particular trial these estimates may be very large. Then at the second level treatment effects for each trial are modelled on one another (equation 3.16), where separation exists there will be outlying points in the regression. The LRF will then be based on the log–likelihood of a model with potentially highly influential outliers. Overall, the work I have conducted suggests that this leads to underestimation of R_{ht}^2, data not shown. R_{ht}^2 estimation will therefore be very unreliable in the presence of separation.

3.6.4 Solution to separation issues

There are various solutions to the issue of separation suggested by Allison (2008). These include: deleting problematic variables; combining categories; reporting only the likelihood ratio statistics; using exact logistic regression; using penalized maximum likelihood or Bayesian estimation.

I could not remove any of the variables and still meet my aims, and I could not report only the likelihood ratio statistics as the surrogate evaluation requires parameter estimates. I could have combined categories but this would have potentially led to loss of information. A Bayesian approach has not been developed for the information theory approach and therefore development of this to overcome separation would have been too involved. Furthermore, Allison (2004) found that generally speaking uninformative priors led to convergence problems. Of the remaining options penalised maximum likelihood was cited as being the best for parameter estimation (Heinze and Schemper, 2002).

The penalized likelihood technique of Firth (1993) was originally introduced to reduce small sample bias in maximum likelihood estimates but has been found to be useful in dealing with separation (Heinze and Schemper, 2002). Firth (1993) found that bias occurred in estimates through trying to derive the score function. The score function is the first derivative of the log-likelihood which is used to obtain maximum likelihood estimates if the function is concave. However, since the score function is unbiased and in the case of separation it is not concave (i.e. has no maxima) problems arise in calculating maximum likelihood estimates. Firth (1993) suggested adding a bias-term to the score function to resolve this.

Heinze and Schemper (2002) applied the technique of Firth (1993) to deal with instances of separation. The bias-term they applied to the score function was based on the information matrix of the parameter that was affected by separation. The influence of the bias term is asymptotically negligible. Firth (1993) showed that using a score function incorporating the bias-term leads to removal of the overall bias in parameter estimation. Heinze and Schemper (2002) assessed this technique and showed that it

was "an ideal solution to separation" as it produces finite parameter estimates that are superior overall to those from alternative methods.

Given that separation causes severe bias in results of surrogacy evaluation I applied the "ideal solution" of Firth (1993) to resolve this issue for the information theory approach in the binary-ordinal setting at the trial level.

3.6.5 Separation: final considerations

There are some final comments on separation to cover:

- In R there are commands `logistf` and `pordlogist` that can be used to apply Firth's approach to both the binary and ordinal cases. There are some bugs in the program for the ordinal case which will be discussed in further detail in the next chapter.
- If a correction of separation is not used, the trials where separation occurs would need to be removed from analysis to avoid the bias previously mentioned. This may lead to large loss of information where trials are set to be small and separation is more likely.
- Firth's approach works best for small samples and in order to calculate the Firth corrected estimate of R^2_{ht} two models for each separate trial will need to be run as opposed to a full model incorporating all trials. As discussed in section 3.4.3 this will return the same estimates as full models but will be more laborious.
- Finally, I should note that, in the case of complete or quasi-complete separation, parameter estimates but not likelihood estimates are affected, hence individual level surrogacy which does not rely on parameter estimation can be estimated without using this alternative method.

3.7 Conclusions: binary-ordinal

I have extended one of the foremost approaches to surrogacy evaluation to the case of a binary surrogate and ordinal true outcome. Information theory can be applied in the multi-trial setting via the LRF. I have provided formulae for applying the LRF at both the trial and individual levels in the binary-ordinal setting and specific modelling approaches that are most appropriate in either case.

At the individual level I adopted a multi-trial LRF approach so that differences between trials can be taken into account, this has not been addressed in previous literature. A multi-trial LRF is not possible at the trial level since a two stage approach is necessary. The fact that R_{ht}^2 cannot account for differences between trials has been identified as an issue by Tibaldi et al. (2003b) who suggested adjusting for trial size as a partial remedy.

I have provided confidence intervals for both the trial and individual levels. Confidence intervals for the individual level offer an improvement on those previously published. I have also adopted the penalized likelihood technique of Firth (1993) as a novel technique for dealing with the common issue of the occurrence of separation in the information theory approach for discrete outcomes.

This work will help researchers assess surrogacy in areas of research where ordinal outcomes are primary outcomes of interest.

Chapter 4. Simulation study: for a binary surrogate and ordinal true outcome

In the previous chapter I outlined how the information theory approach to surrogate evaluation has been extended to the case of a binary surrogate and ordinal true outcome. It was important that this extension was well investigated through simulation study and case study to determine how well it works. This chapter will outline the approaches and results of the simulation study that I have performed in order to achieve this.

This chapter outlines: the simulation study process (the models required to set it up and practicalities of doing so) in Section 4.1.1; the results of the simulation at the individual and trial level, sections 4.2.1 and 4.2.2; and finally give my conclusions.

As in previous chapters: S denotes a surrogate; T a true outcome; Z was treatment; i=1,..,N trials and j=1,..,n_i patients per trial.

4.1.1 Simulation study: set up

The simulation study determines how well the information theory surrogate evaluation approach in the binary-ordinal setting performs under a variety of different scenarios.

In the simulation study continuous variables for S and T were first simulated, these were dichotomised or split into categories to represent a binary S and an ordinal T, as in Pryseley et al. (2007). Pryseley et al. (2007) highlighted that R_h^2 represents surrogacy for the latent unobserved continuous surrogate and true outcomes. This means that surrogacy is established in the continuous underlying setting. Categorisation of the variables leads to lower strengths of surrogacy in the observed discrete setting. However, at the trial level the "relationship between the treatment effects on the latent-continuous and observed-binary surrogate endpoints was linear", Pryseley et al. (2007). Hence, the value of R_{ht}^2 is valid on both scales and the value of R_{ht}^2 should theoretically be the same in the underlying and observed settings.

It was valid to simulate continuous outcomes which were then categorised due to the precedent set in other publications. Furthermore, categorised continuous variables are true to life, since treatment measures often represent underlying unknown continua. This will be discussed in more detail in Section 4.1.3.3.1.

In order to determine the best way of setting up the simulation study I investigated previous methods used for surrogate evaluation of non-causal approaches, i.e. not principal stratification (see section 2.3).

In this section I:

- draw conclusions on the best simulation method in the binary-ordinal case, section 4.1.1.4;
- outline the models required to implement this, section 4.1.2;
- and discuss some practicalities:
    - general practicalities, section 4.1.3.1;
    - theoretical and coding practicalities, section 4.1.3.2;
    - and finally some more in-depth considerations, see section 4.1.3.3.

4.1.1.1 Set up: previous methods

Simulation methods not considered were those under direct and indirect effects or principal stratification. These methods simulate counterfactual results and relationships which are not necessary for the information theoretic approach. Hence, these approaches were not suitable. There are two simulation methods so far published in the surrogate evaluation literature for non-counterfactual approaches. I in this section I discuss the set-up of each alongside criticisms and outline how these can be implemented.

4.1.1.2 Previous methods: simulation method one

The first method to setting up a simulation study found in the surrogate evaluation literature was a well-established one by Burzykowski et al. (2005) which was duplicated in a number of publications by the same group of authors. These authors created simulation studies where the true values of R_h^2 and R_{ht}^2 were known. The amount of variation at individual and trial level, trial size and number of trials were

varied to see if this influenced the ability of their method to estimate the correct surrogacy strength.

The model seen in equation 4.1 is based on the meta-analytical approach, see section 2.2.1. The parameters of the model are given set values to enable simulation.

$$S_{ij} = 0.50 + m_{S_i} + (0.05 + \alpha_i)Z_{ij} + \varepsilon_{S_{ij}},$$

$$T_{ij} = 0.45 + m_{T_i} + (0.03 + \beta_i)Z_{ij} + \varepsilon_{T_{ij}},$$

$$(m_{S_i}, m_{T_i}, \alpha_i, \beta_i) \sim \mathrm{N}(0,\mathrm{D}) \ , \left(\varepsilon_{S_{ij}}, \varepsilon_{T_{ij}}\right) \sim \mathrm{N}(0, \sum), \qquad \textbf{\textit{4.1}}$$

$$D = \delta_C^2 \begin{pmatrix} 1 & 0.85 & 0 & 0 \\ 0.85 & 1 & 0 & 0 \\ 0 & 0 & 1 & \rho \\ 0 & 0 & \rho & 1 \end{pmatrix} \rho^2 = 0.90,$$

$$\sum = \delta^2 \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$$

This is a joint mixed effects model of S and T regressed on treatment, with individual errors distributed N(0, $\sum$), and random intercepts and random treatment effects for trial distributed N(0,D). Here, δ^2 was set to either 0.1 or 3 and δ_C^2 to 0.1 or 10 these denote the within trial and between trial variability respectively Burzykowski et al. (2005).

In the simulation the fixed effects coefficients of the meta-analytical model, in equation 4.1, were given proposed values, the fixed intercepts were 0.50 and 0.45 and fixed treatment effects 0.05 and 0.03. The covariance parameters of the error terms and the jointly distributed random effects were selected in matrices $\sum$ and D in order to create certain strengths of surrogacy. When the covariation of the random treatment effects of S and T was ρ^2=0.90 in D surrogacy at the trial level was 0.90. Surrogacy can readily be set to some other value via this component of D. Surrogacy at the individual level was set via the covariation of S and T in $\sum$ which in this example was 0.80. Each random variable in the joint model can be simulated using multivariate normal distributions based on the covariance matrices $\sum$ and D. Putting

all of these components together using the above formulae means continuous values for S and T can be simulated for each patient. In our case the continuous S and T could then be categorised to create a binary surrogate and ordinal true outcome.

The true values of R_h^2 and R_{ht}^2 can be set with precision using this simulation method. This allows assessment of the amount of bias in estimates for each scenario.

The meta-analytical approach was the predecessor to the information theory approach and analogous in many respects. It is a highly valued approach to surrogacy evaluation in its own right and in the continuous setting has few drawbacks. Therefore, it is valid to simulate surrogacy using the models of the meta-analytical approach for use in assessing the measures of the information theory approach.

4.1.1.3 Previous methods: simulation method two

The second simulation method considered, proposed by Lassere et al. (2007a), was conceptually appealing and compared various surrogate evaluation approaches to determine the best. They set up their simulation to infer certain treatment relationships between the surrogate and true outcome. They then overlaid a corresponding correlation between the surrogate and true outcome at the individual level. In surrogacy evaluation it is important to take treatment effects into account as reliance on a correlation between surrogate and true outcome without reference to treatment is insufficient. Therefore, any measure that validates a surrogate based on correlation regardless of the treatment effect relationships between S and T is a poor means of determining surrogacy. This simulation approach would highlight those surrogacy evaluation approaches that fail in this regard.

However, a detailed description of the method of the simulation was not provided in the Lassere et al. (2007a) paper. Therefore, a formal description of the method cannot be provided and only generalities can be discussed.

This simulation method investigated various magnitudes of correlation of S and T combined with different strengths of treatment effect agreement between S and T. The true strength of surrogacy was determined by the treatment effect relationships. So that in some situations the surrogacy message from the correlations did not agree

with true strength of surrogacy. A good surrogacy measure was one that could determine the true strength of surrogacy and not rely on the magnitude of correlation.

The scenarios proposed by Lassere et al. (2007a) were:

- Strong positive correlation between S and T at the individual level.
- Weak positive correlation between S and T at the individual level.
- No correlation between S and T at the individual level.

Within each of these there were the additional four scenarios of:

- Treatment effect (Z) on T and not on S.
- Treatment effect on S and not on T.
- No treatment effects.
- Or treatment effects on both.

Trying to emulate this approach raised several issues and led to two main arguments against its use:

1. The Lassere et al. (2007a) method does not specify the exact value of R_h^2 and R_{ht}^2. Unlike Lassere et al. (2007a) my simulation was intended to only investigate one approach to surrogacy evaluation which would provide no reference point for comparison. Scenarios where there were "Treatment effects on both" represent the case of a strong surrogate. But what strength of surrogacy does this represent, say 0.70 or 0.90? And if our surrogate evaluation measures returned a value of 0.68 would this be a bias of 0.02 or 0.22? Therefore, this simulation method was not adequately framed to determine the degree of precision of our surrogate evaluation measure.
2. There are severe difficulties in modelling this situation given the complicated nature of the relationships under investigation.

4.1.1.4 Previous methods: conclusions

Attempting implementation of the Lassere et al. (2007a) method highlights certain practical and conceptual difficulties. The most striking being the inability of the method to tell how biased a surrogacy evaluation method is compared to the real strength of surrogacy. Since this was an integral part of what I wished to achieve I adopted the structural set up of Burzykowski et al. (2005). However I was persuaded

to incorporate some aspects of the concepts of the Lassere et al. (2007a) method by providing a wider investigation of the influence of different and conflicting strengths of surrogacy at the individual and trial levels.

The main aims of the simulation study were therefore to investigate the impact on R_h^2 and R_{ht}^2 of:

- higher and lower values of the true strength surrogacy in R_h^2 and R_{ht}^2;
- strength of surrogacy disagreeing between the trial and individual levels;
- non-proportional odds in the ordinal true outcome;
- and varying the number of trials and the number of patients per trial.

4.1.2 Set up: binary-ordinal setting

First I will describe the scenarios investigated in the simulation, the exact working or set-up of these scenarios is described in what follows or in section 4.1.3. Then I will describe the mathematical models that are used to simulate the data. Each component of the model is simulated or set and then S_{ij} and T_{ij} calculated on the basis of these as normally distributed variables that are then categorised as described in 4.1.3. Finally, when all the parameters have been discussed a model with the set parameter values is given.

The scenarios investigated in the simulation for the binary-ordinal setting are shown in Table 4.1. The number and size of trials were chosen to correspond to the previous work of Burzykowski et al. (2005). Strengths of surrogacy were set to be high or low or disagree in strength at trial and individual levels, this will be discussed further below. Assessment of the impact of non-proportional odds was also conducted in the simulation. The set-up of the non-proportional odds setting will be discussed further in 4.1.3. I will now outline the simulation mathematically.

Factor varied under simulation study	**Levels of factor**
Number of trials	5, 10, 20 or 30
Number of patients per trial Small trial size Large trial size	 10, 20, 40 or 60 100, 150, 200 or 300
True surrogacy strength	R_h^2=0.64 or 0.30 and R_{ht}^2=0.90 or 0.30
Agreement of surrogacy strength	
• Agree both strong or both weak	R_h^2= 0.64 & R_{ht}^2=0.90 Or R_h^2= 0.30 & R_{ht}^2=0.30
• Disagree one weak one strong	R_h^2= 0.64 & R_{ht}^2=0.30 Or R_h^2= 0.30 & R_{ht}^2=0.90
Trial level: Adherence to the proportional odds assumption	 Proportional odds Non proportional odds

Table 4.1: *Scenarios investigated in this binary-ordinal simulation*

I based my simulation on the model seen in equation 4.2, to produce variables Z_{ij}, S_{ij} and T_{ij}, this was advocated by Burzykowski et al. (2005). The model in 4.2 is a joint mixed effects model of S_{ij} and T_{ij} regressed on treatment, with individual errors distributed N(0, $\sum$), and random intercepts and random treatment effects for trial distributed N(0,D). The parameters of the model μ_S, μ_t, a, b, ρ^2, ψ^2 ,$\delta^2$and δ_c^2 which will be discussed below are given set values to enable simulation.

$$S_{ij} = \mu_S + m_{S_i} + (a + \alpha_i)Z_{ij} + \varepsilon_{S_{ij}},$$

$$T_{ij} = \mu_T + m_{T_i} + (b + \beta_i)Z_{ij} + \varepsilon_{T_{ij}},$$

$$(m_{S_i}, m_{T_i}, \alpha_i, \beta_i) \sim \text{N}(0,\text{D}) \ , \left(\varepsilon_{S_{ij}}, \varepsilon_{T_{ij}}\right) \sim \text{N}(0, \sum\ ), \qquad \textbf{4.2}$$

$$D = \delta_C^2 \begin{pmatrix} 1 & 075 & 0 & 0 \\ 0.75 & 1 & 0 & 0 \\ 0 & 0 & 1 & \rho \\ 0 & 0 & \rho & 1 \end{pmatrix} \rho^2 = 0.90,$$

$$\sum = \delta^2 \begin{pmatrix} 1 & \psi \\ \psi & 1 \end{pmatrix}$$

μ_S, μ_t, a, b, $\delta^2$and δ_c^2 are set to correspond to the simulation of Burzykowski et al. (2005), Tilahun et al. (2008b) and Pryseley et al. (2007). The fixed intercepts, m_{S_i} and m_{T_i} of equation 4.2, were set to 0.50 and 0.45 respectively, and the fixed treatment effects, , a and b of equation 4.2, to 0.05 and 0.03. δ^2 denotes the trial level variability and δ_C^2 the individual level variability these were both set to 3, see Burzykowski et al. (2005).

The components of the covariance matrices were set to infer certain strengths of surrogacy with ρ^2 setting trial level surrogacy and ψ^2 individual level surrogacy, and correspond to the scenarios in Table 4.1. To simulate weak surrogacy at respective levels I set ρ^2 =ψ^2=0.30, and for strong surrogacy at respective levels I set ρ^2 =0.90 and ψ^2=0.64. As can be seen in Table 4.1 different combinations of these strengths of surrogacy are investigated in this simulation.

The values of the fixed trial intercepts and treatment effects differ to those of Burzykowski et al. (2005) only in scale. The selected values were identical to those adopted by this group of authors when they published simulation studies for information theory measures of surrogacy (Tilahun et al., 2008b) and (Pryseley et al., 2007). When R_h^2=0.64 and R_{ht}^2=0.90 our results will be roughly comparable to these publications. The only difference will be in the choice of lower co-variation in the random intercepts (0.75 instead of 0.85), see variance covariance matrix D in equation 4.2. This value was chosen to be more in keeping with all values of R_{ht}^2 in my simulation scenarios, which was also done by Abrahantes et al. (2004). This did not change the strength of surrogacy but 0.75 was closer to the other component, ρ in D, when surrogacy was weak, ρ =0.55 (i.e. $\rho^2 = R_h^2 = 0.30$), and in using this term the components of D were not too different to each other which may be more realistic to real life scenarios. For the simulation set up with a summary of all of these chosen parameter values see Table 4.3.

$$S_{ij} = 0.50 + m_{S_i} + (0.05 + \alpha_i)Z_{ij} + \varepsilon_{S_{ij}}, \qquad 4.3$$

$$T_{ij} = 0.45 + m_{T_i} + (0.03 + \beta_i)Z_{ij} + \varepsilon_{T_{ij}},$$

$$(m_{S_i}, m_{T_i}, \alpha_i, \beta_i) \sim \mathrm{N}(0,\mathrm{D}) \ , \left(\varepsilon_{S_{ij}}, \varepsilon_{T_{ij}}\right) \sim \mathrm{N}(0, \sum),$$

$$D = 3\begin{pmatrix} 1 & 075 & 0 & 0 \\ 0.75 & 1 & 0 & 0 \\ 0 & 0 & 1 & \rho \\ 0 & 0 & \rho & 1 \end{pmatrix}, \rho^2 = \mathrm{R}_{\mathrm{ht}}^2 = 0.90 \text{ or } 0.30,$$

$$\sum = 3\begin{pmatrix} 1 & \psi \\ \psi & 1 \end{pmatrix}, \psi^2 = R_h^2 = 0.64 \; or \; 0.30$$

The models described above can be easily implemented, although several practical aspects of performing this simulation study have to be discussed.

4.1.3 Set up binary-ordinal: practicalities

In this section I outline and justify the setup of this simulation study. I start with general practical considerations and then move on to theoretical and coding issues. I finish by outlining some more involved issues.

4.1.3.1 Practicalities: general

- A starting seed was selected and the full code used to run the simulation was saved and backed up.
- All datasets simulated represent one iteration of a scenario and contain a set number of trials of a given size. Each trial dataset was stored as an RDS file. The variables within each dataset were; the continuous true and surrogate outcomes, the binary surrogate and ordinal true outcomes, the number of trials and treatment group.
- For each dataset simulated a surrogate analysis was performed. The surrogate analysis results for every dataset simulated with a record of the occurrence of any modelling issues was recorded.
- There were 250 simulated datasets created for each scenario. The median value of R_h^2 and R_{ht}^2 will be reported in line with Pryseley et al. (2007). This was appropriate for our results as some distributions of simulated R_h^2 and R_{ht}^2 values were skewed especially where the true value was high. The median R^2 value for a certain scenario served as a measure of the bias present in our results. In addition, I present the interquartile ranges (IQRs) of the 250 R_h^2 and R_{ht}^2 for each scenario. I give the IQRs as one value of the length of the interquartile range rather than the upper and lower bounds to simplify tables and ease comparisons across scenarios where comparisons of the degree of precision in estimation need to be assessed. These ranges will give an indication of the degree of precision of these estimates under each condition. Finally, the coverage and median upper and lower bounds of the confidence intervals will be presented. The coverage of the CIs, at the trial level, is calculated as the % of intervals in a particular scenario that include the true value of surrogacy as set by the simulation. However, this will be based on the empirical asymptotic truth at the individual level, as will be discussed

further in section 4.1.3.3.3.1. The confidence intervals were calculated using the noncentral χ^2 distribution as discussed in section 3.5. Presenting these statistics will inform on how well these intervals performed in the simulation study.

- Burton et al. (2006) recommend recording and appropriately dealing with the occurrence of errors in simulation studies. I recorded convergence issues using the R command `tryCatch()`. This function allows warning and error messages to be reported without breaking the loop. If the models failed for any reason this was recorded.

4.1.3.2 Practicalities: theoretical and coding

- I set the amount of between and within trial variation to the same value, $\delta^2 = \delta_C^2 = 3$ (see Equation4.2). This was appropriate as Abrahantes et al. (2004) showed that bias was lower when variation was at least as large at the lower hierarchical level than at the higher one. Every publication by this group of authors since Abrahantes et al. (2004) has used $\delta^2 = \delta_C^2 = 3$.
- An investigation of simulation studies with a three level hierarchy, individual patients, centres and trials, have been investigated in Burzykowski et al. (2005). For this study I was only interested in two levels, individual patients and centres within a trial. This was because these two levels of the hierarchy corresponded to the case study presented in the next chapter.
- The work of Tilahun et al. (2007) suggested that coding treatment group in the simulation study (1,-1) was better than (0,1). The latter meant that the variance of one treatment arm was likely to be less than that in the other which would only be relevant in very particular studies.

  In my study coding the treatment (1, -1) led to strong instances of separation of treatment groups within trials. Often, perfect separation would occur and patients on treatment 1 would all have an outcome but no patients on -1 would, even for very large trial sizes. This seemed like an unrealistic scenario. As previously discussed, the occurrence of separation led to serious estimation bias, see section 3.6. Furthermore, it seemed unsound to let the coding of treatment influence results when it was not specified as part of the simulation model. Hence, a binary treatment variable was coded (0.5,-0.5) for

simulation purposes. Under this coding the number of separation issues was greatly reduced. This coding was still in keeping with the proposals of Tilahun et al. (2007) as it still allows for equal variance in both treatment arms.

- In order to create a binary surrogate outcome, the simulated continuous surrogate was dichotomised at the mean. This was in keeping with previous publications in information theory and the meta-analytical approach (Tilahun et al., 2008b), (Pryseley et al., 2007) and (Burzykowski et al., 2005).
- I simulated a seven category true outcome to match the Oxford Handicap Scale, (Van Swieten et al., 1988), used in the case study. This will be discussed in the next chapter.
- Continuous simulated true outcome variables were created based on a set treatment relationship using linear models. To create ordinal outcomes with proportional odds the continuous variables were categorised using six evenly spaced cut off points. The cut off points were determined according to the quantiles of the true outcome variable, see column 1, Table 4.2. Since the continuous variables are linear and cut points are the same across treatment groups the same relationship between the two groups should be retained within each resultant category. Hence, the odds should be proportional across categories of the ordinal variable.
    - The cut points of the ordinal variable were chosen to be equally spaced to make sure that no one category was more likely to suffer missing data than another. I anticipate no issues of generalizability of the findings of this scenario to proportional odds scenarios were this is not the case.
    - An illustrative simulation was run to see the true simulated odds ratios for a particular scenario. This simulation was run 1000 times for the case of 30 trials and 300 patients, and strong surrogacy at both levels, to provide an example of these values. The odds ratios at each cut point were calculated for each simulated ordinal true outcome and the median odds ratios are shown in Table 4.2. As can be seen the odds ratios for the proportions odds setting are very similar for each cut

point, hence we can conclude that in this simulation the proportional odds scenario holds true to its name.

	Column 1: Proportional odds			Column 2: Non proportional odds		
				Coding for divergent treatment group		
Ordinal category	Quantiles continuous	Quantiles ordinal	Simulated OR	Quantiles continuous	Quantiles ordinal	Simulated OR
1	≤ 0.143	0.143	0.9714	≤ 0.25	0.25	0.4900
2	0.143-0.286	0.143	0.9774	0.25-0.286	0.036	0.9806
3	0.286-0.429	0.143	0.9776	0.286-0.429	0.143	0.9767
4	0.429-0.571	0.143	0.9731	0.429-0.571	0.143	0.9845
5	0.571-0.714	0.143	0.9838	0.571-0.714	0.143	0.9827
6	0.714-0.857	0.143	0.9714	0.714-0.857	0.143	0.9749
7	≥ 0.857	0.143	-	≥ 0.857	0.143	-

Table 4.2: *Categorisation of continuous true outcome into ordinal true outcome. Simulated odd ratios were based on 1000 runs of the study set up with 30 trials, 300 patients, and strong surrogacy at both levels, the median odds ratios over all simulated cases are given for each cut point.*

- It was important in the case of simulating non-proportional odds variables that the treatment relationships of the underlying continuous outcomes were not altered via categorisation. Therefore, for the non-proportional odds scenario one treatment was categorised according to equal quartile cut-points, as in column 1 Table 4.2. The divergent treatment group was also categorized based on these equal cut points except for two categories. Category one of the divergent treatment group incorporated a much larger amount of quantiles, the first 0.25 quantiles, meaning category 2 had much fewer quantiles, 0.036 quantiles (see Table 4.2). In the divergent treatment group, the odds in

relation to the other treatment group are not the same as each other or to the other categories on the scale. Hence, the odds are not proportional. The remaining categories retain the treatment relationships set by the simulation and have proportional odds in relation to each other. Refer again to the median odds ratios taken from the simulation of true ordinal outcomes in Table 4.2, this time for the case of non-proportional odds. We see that the odds ratios are very similar for all but the first cut of the ordinal true outcome. Here the odds ratio is 0.49 as opposed to approximately 0.98 for the other cut points. Therefore, this shows that as anticipated the odds ratios in the simulation are proportional for all but one cut point of the ordinal true outcome. Hence, an ordinal true outcome with non-proportional odds was simulated without fundamentally changing the treatment relationships simulated at the underlying continuous-continuous setting.

- This simulation method was adapted from the simulation study of ordinal outcomes of McHugh et al. (2010b). In that study it was hypothesised that a treatment only benefited one category of the simulated ordinal outcome, affecting the resultant proportional odds assumption. In keeping with this work I hypothesized that, in addition to the treatment relationship simulated in the underlying continuous setting, the treatment resulted in some patients who would otherwise have had a category two outcome to have a category one outcome.

4.1.3.3 Practicalities: in-depth issues

Several additional avenues of investigation were identified after further reading and investigation of preliminary results. These were: how to deal with issues of loss of information; a technique for dealing with separation of treatment groups; and finally the choose of R commands to be used in the simulation.

4.1.3.3.1 Loss of information

I first simulated a continuous surrogate and a continuous true outcome, these were then dichotomised or categorised into binary and ordinal outcomes. Tilahun et al. (2008b) and Pryseley et al. (2007) investigated surrogacy in the binary-binary and

binary-continuous settings respectively. They demonstrated that dichotomisation or categorisation of continuous outcomes leads to estimates of R_h^2 that are much reduced compared with the surrogacy strength set at the latent continuous setting.

4.1.3.3.2 Loss of information: Simulation rationale

Dichotomised or categorised surrogates are less informative than continuous surrogates because they provide less information. However categorised versions of continuous outcomes are present in real life scenarios and are likely to suffer the same issue, Tilahun et al. (2008a). The ordinal measure Oxford Handicap Scale (OHS) is a reliable seven category measure of death and disability. This measure has less inter-observer variability than might be present in a measure of death and disability say on a 100-point scale. However, the true underlying amount of disability is likely to be on a more finely graded scale than the OHS. If it were possible to reliably measure the true amount of disability on a continuous scale this would be much more informative than the OHS. In the absence of a reliable continuous measure, OHS is a good approximation of this underlying continuum. Therefore, simulating continuous variables that are then categorised is a good way of providing data that represents this real life scenario.

4.1.3.3.3 Loss of information: theoretical impact

The selected strength of R_h^2 in the simulation represents the underlying latent continuous variables but the R_h^2 value of the observed binary and ordinal variables was unknown. Several publications investigated the bias imposed by dichotomising and categorising variables, (Taylor and Yu, 2002, Cochran, 1968, Bollen and Barb, 1981, Krieg, 1999). Cox (1957) showed that the maximum amount of information retained when a dichotomisation occurred was 63.7%. This retention of information increased as the number of groups created via categorisation increased. The amount of information retained for the largest categorisation reported of six groups was 94.20%: my ordinal variable has seven groups.

It would be useful to be able to determine the true strength of surrogacy in the observed binary-ordinal case where this was set to a certain strength at the underlying continua. It was possible to investigate the impact on surrogacy of

dichotomising continuous outcomes under the meta-analytical approach but not under the information theory approach. Also, it was not possible to do this for ordinal outcomes under either framework: we therefore do not know the true surrogacy strength in the observed binary-ordinal setting.

I have shown that the estimates of R_h^2 in this simulation study in the observed binary-ordinal setting will be substantially lower than in the underlying latent setting. I have therefore included a third scenario in the simulation study, where R_h^2=1 in the latent continuous setting. This was done to establish what the ceiling effect was for R_h^2 in the binary-ordinal setting. The results of this scenario will be discussed later, section 4.2.1.5.

4.1.3.3.3.1 Coverage of confidence intervals

Given that the true value of R_h^2 is not known in order to calculate the coverage of the confidence intervals the large scale approximation of R_h^2 as determined by the simulation for each scenario will be used instead.

4.1.3.3.4 Separation

The large bias imposed through the occurrence of complete and quasi-complete separation was discussed in section 3.6. In this case, separation occurred in regression analysis when zero cells were present in trial cross-tabulations of treatment and surrogate or treatment and true outcome.

4.1.3.3.4.1 Separation at the individual level

In section 3.6.3 I discussed how treatment effect estimates can be severely biased in the presence of separation. Treatment effect estimates are crucial for calculating trial level surrogacy. It was also noted that this issue does not impact individual level surrogacy, as this does not rely on treatment effect estimation. It only uses log-likelihood statistics which are not affected. However, there is one instance of the presence of zeros which does impact on the individual level surrogacy. If patients in a certain trial, regardless of treatment, do not have any surrogate outcomes or all have surrogate outcomes the models failed, since the surrogate variable was effectively only returning one value. Hence the first model of true outcome regressed on treatment worked, but adding the surrogate as an explanatory variable in the

second model led to model failure. If you recall I was interested in the amount of information on the true outcome that was provided by the surrogate and calculated this by using the -2*log-likelihood difference between the two models described (with and without adjustment for the surrogate), see section 3.3.2. If the second model does not converge the log-likelihood could not be calculated for this model.

There were two ways I could have dealt with this issue. One was to remove all the trials where separation of this kind occurred (occurrence can be as high as 17% of trials). Secondly, since the surrogate does not provide any information on the true outcome effectively the log likelihood of the second model can be thought of as just as informative as the first. In other words, the surrogate provides no additional information on the true outcome. I believed the former technique was inappropriate as it was throwing out information on trials where the surrogate was uninformative and therefore may lead to inflated estimates.

The results of both techniques showed that indeed this was the case (data not shown). Where instances of this pattern of zeros were prevalent the results showed that the trial removal technique led to overestimation compared to the alternative technique where estimation was very good. Therefore, where no surrogate outcome was recorded for a particular trial in the analysis the difference in the log-likelihood between the two models with and without adjustment for the surrogate was recorded as zero.

4.1.3.3.4.2 Separation at the trial level

As previously mentioned, separation has the most severe impact on trial level surrogacy where stage one estimation was badly impacted by separation. These estimates were regressed at the second stage of modelling and could result in influential outlying points, see section 3.6.3. To deal with issues of separation I proposed the use of the penalised likelihood technique of Firth, see section 3.6.4. This technique allowed trials where separation occurred to be retained in the analysis and provided sensible treatment effect estimates. In the simulation, the penalized likelihood technique results will be compared to a trial removal technique where

trials where separation occurs were removed from analysis (resulting in loss of information).

There are two points that need to be discussed in relation to separation in terms of the setup of this simulation study. Firstly, the method of implementing the technique of Firth in R will be discussed in Section 4.1.3.3.5. Secondly, there was an issue in calculating R^2_{ht} under the trial removal technique if a large number of trials with separation in a particular simulated dataset were removed from analysis.

Scenarios where there were only ten patients per trials suffered to a much larger extent from issues of separation at the trial level compared to other scenarios. However, the high instances of separation in these smaller trial size scenarios meant that often less than three trials were available within one scenario for the calculation of R^2_{ht}. This was the case regardless of the number of trials. I set the analysis so that at least three trials had to be available for a particular simulated scenario. This was done, because second stage models were based on the treatment effect estimates for each trial, and these models do not converge with less than three data points. This meant that simulated datasets where less than three trials were available would not have results returned for R^2_{ht}. The simulation ran until 250 simulated datasets were created for each scenario where results were possible. The number of failures were recorded.

```
Call:
lm(formula = bi ~ mui + ai)

Residuals:
ALL 3 residuals are 0: no residual degrees of freedom!

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.6306         NA      NA       NA
mui           2.8839         NA      NA       NA
ai            1.3637         NA      NA       NA

Residual standard error: NaN on 0 degrees of freedom
Multiple R-squared:     1,      Adjusted R-squared:   NaN
F-statistic:   NaN on 2 and 0 DF,  p-value: NA
```

Figure 4.1: *R output for second stage trial level models with only three trials*

Furthermore, where there were only three trials at the second stage of modelling the calculation of R_{ht}^2 was complicated due to an issue with lack of degrees of freedom. The output from such a situation can be seen in Figure 4.1. Such a model returns a log-likelihood of infinity and hence a R_{ht}^2 of 1; an erroneous result. Where there was an issue with degrees of freedom the calculation of R_{ht}^2 was not performed.

4.1.3.3.5 Modelling in R

In this section I briefly describe the logic behind the R functions chosen to perform generalised linear and cumulative logit link regression for the binary surrogate and ordinal true outcome respectively. I will also discuss the penalised likelihood functions needed in the binary-ordinal setting to deal with instances of separation.

Cumulative logit link proportional odds models were used to analyse the ordinal true outcome as a response variable in the models required at both the trial and individual levels. Alternative link functions were not considered.

The R functions `lrm` from the package RMS and the `polr` from the package MASS both have a proportional odds set up. `lrm` was based on equation 7.6 and `polr` takes a latent variable method described in section 7.2.3 of Agresti (2014). Thompson (2009) provided an R manual to accompany Agresti (2002) which has been endorsed by Agresti (2012). Thompson (2009) recommend the use of `lrm` and `polr` as being transparent and easily fitted cumulative logit models. The `lrm` function was used to calculate R_h^2 at the individual level and `polr` for R_{ht}^2 at the trial level.

For models with the binary true outcome as a response variable the usual generalised linear model command `glm` was used, for generalised linear models. Second stage models were modelled using the command `lm`, for ordinary linear regression.

4.1.3.3.5.1 Modelling the penalized likelihood technique

To deal with issues of separation I suggested using the penalised likelihood technique of Firth (1993), see section 3.6.4. For generalised linear models a command in R called `logistf` from the package of the same name can deal with separation. This package was based on the work of Heinze and Schemper (2002) and

Firth (1993). `logistf` performs better where datasets are small and uncomplicated by hierarchies. Therefore, single models for each trial within a dataset were computed as opposed to a full model incorporating all trials (see 3.4.3). This single model method returned the treatment effect estimates of each trial as required. Where separation does not occur the treatment effect estimates that were required were calculated using the `glm` command, otherwise the `logistf` command was applied.

There is also a command available which provides a penalized likelihood approach for cumulative logit link models called `pordlogist` from the package OrdinalLogisticBioplot (Hern'andez, 2013). This command uses a simpler penalized likelihood technique in line with Firth (1998) based on Le Cessie and Van Houwelingen (1992). This is an auxiliary command as part of the OrdinalLogisticBioplot package. This command has some bugs. Where zero cells were present in both treatment arms, i.e. an empty category, in a certain trial the command failed. On the advice of Vicente Villardón (2015), one of the authors of the package, I recoded the ordinal variable to ignore zero categories which resolved this issue. In the ordinal models used the odds were considered as proportional between categories meaning that the removal of zero categories should not change results. The `polr` function was used to estimate treatment effects for trials where separation does not occur and `pordlogist` when it did. Using both `logistf` and `pordlogist` means that all trials were retained in the simulation study despite the presence of separation and so the bias imposed by this issue was removed.

4.1.4 Set up: Conclusions

I have investigated historic simulation studies in the surrogate evaluation context and concluded that a method based on a joint mixed model was best for my purposes. This method permits the exact strength of surrogacy to be set (at the individual level this is only at the underlying continuum) to allow the degree of bias and variability in results to be assessed.

The simulation will investigate surrogacy: for varying numbers and sizes of trials; for non-proportional odds; and when surrogacy strength differs at the trial and individual levels. A ceiling effect for the individual level will be investigated by simulating a

perfect surrogacy scenario. A solution for separation in the trial level models will be investigated using a penalized likelihood technique and compared to a trial removal technique.

4.2 Simulation study: Results

For a full description of the scenarios investigated see Table 4.1 and section 4.1.4.

The summary results given are: the median R^2 for a given scenario; the variation of R^2; and the median upper and lower bounds of all of the confidence intervals calculated for each R^2 for a particular setting. The tables in the remainder of this chapter show results for a subset of the trial sizes investigated. Unless otherwise stated, results for scenarios not included in the tables were consistent with those shown. Full tables for every setting are provided in Appendix A.

Some of the results show conflicting biases especially at the trial level. Therefore, the results can be difficult to interpret. To aid this Table 4.3, for reference, outlines the relevant issues and what scenarios they affect and also in which section they are described.

	Issue	Scenarios affected	Reason	Discussed in section(s):
R_h^2	R_h^2 estimates substantially lower than true value in continuous setting	All scenarios	Loss of information due to categorisation	4.1.3.3.1, 4.2.1.2 and 4.2.1.5
R_{ht}^2	Underestimation	Worse for: • Larger numbers of trials • Strong surrogacy • Small trial sizes	Inefficiency in estimation due to categorisation and the use of a two stage approach	4.2.2.1.2
	Overestimation	Worse for: • Small numbers of trials • Weak surrogacy	Model fitting issues	4.2.2.2.5
	Separation	All scenarios	Zero cells in trial crosstabs resolved through use of Firth (1993) technique	4.2.2.5 and 4.2.2.6

Table 4.3: *Issues present in results of simulation study*

I will first discuss the results for individual level surrogacy R_h^2 in section 4.2.1 and then those for trial level surrogacy R_{ht}^2 in section 4.2.2.

4.2.1 Results: individual level surrogacy

In this section, I describe:

- the scenario where surrogacy was set to be strong at both the trial and individual levels;
- where this was weak at both levels;
- where the strength of surrogacy differs at trial and individual levels;

- the behaviour of R_h^2 under deviations from the proportional odds assumption;
- the ceiling investigation results;
- and finally compare the results to those of previous publications.

4.2.1.1 Individual level surrogacy R^2_h: strong surrogacy

Here surrogacy was set to be strong at both levels, R_h^2=0.64 and R_{ht}^2=0.90, in the underlying continuous setting. Due to loss of information the observed results for the binary-ordinal setting were expected to be lower than in the underlying continuum.

As can be seen in Table 4.4 surrogacy strength for the observed binary-ordinal setting was much lower than that set in the underlying continuous setting (R_{ht}^2 =0.64). As trial sizes increased the value of R_h^2 converged to around 0.29, suggesting that this approximates the true surrogacy strength for the observed binary-ordinal setting. Small number and sizes of trials scenarios also return results consistent with 0.29.

The IQRs of R_h^2 were fairly narrow for small numbers and sizes of trials but this decreased further as the size and number of trials increased. The coverage of nearly all scenarios was 100% indicating that the intervals are conservative (the only exceptions were for smaller numbers of trials with larger trial sizes). However, the median confidence intervals had sensible ranges. The width of these median intervals decreased with an increase in the size of trials, with good precision for larger trial sizes. However, increases in the number of trials had little impact on the intervals.

Number of trials	Trial size	R_h^2	IQR R_h^2	% cover CIs	Median CI lower	Median CI upper
5	10	0.347	0.154	100%	0.046	0.757
5	60	0.305	0.086	100%	0.126	0.519
5	100	0.300	0.063	99%	0.155	0.465
5	300	0.304	0.058	96%	0.213	0.401
10	10	0.337	0.107	100%	0.046	0.742
10	60	0.293	0.063	100%	0.121	0.499
10	100	0.302	0.056	100%	0.156	0.467
10	300	0.294	0.040	99%	0.207	0.389
20	10	0.342	0.072	100%	0.046	0.743
20	60	0.297	0.039	100%	0.121	0.501
20	100	0.293	0.036	100%	0.151	0.455
20	300	0.292	0.031	100%	0.205	0.386
30	10	0.340	0.071	100%	0.047	0.739
30	60	0.295	0.033	100%	0.121	0.500
30	100	0.294	0.027	100%	0.151	0.454
30	300	0.293	0.025	100%	0.206	0.387

Table 4.4: *Simulation study: Median R^2_h estimates based on 250 simulations for each scenario, where true values set to: R^2_h =0.64 and R^2_{ht}=0.90.*

4.2.1.2 Individual level surrogacy R^2_h: weak surrogacy

The findings for strong surrogacy were mirrored when surrogacy was set to be weak at both the trial and individual level; R^2_{ht}=R^2_h =0.30. The expected strength of surrogacy in the unobserved latent setting was 0.30 however the estimated value of R^2_h for the observed binary surrogate converges to around 0.13, see Table 4.5.

Number of trials	Trial size	R_h^2	IQR R_h^2	% cover CIs	lower 95% CI	upper 95% CI
5	10	0.216	0.142	100%	0.013	0.659
5	60	0.147	0.056	100%	0.030	0.336
5	100	0.135	0.043	100%	0.038	0.282
5	300	0.135	0.033	99%	0.070	0.216
10	10	0.211	0.101	100%	0.018	0.648
10	60	0.141	0.037	100%	0.029	0.327
10	100	0.136	0.032	100%	0.041	0.281
10	300	0.134	0.023	100%	0.070	0.214
20	10	0.212	0.070	100%	0.018	0.645
20	60	0.139	0.029	100%	0.030	0.325
20	100	0.136	0.021	100%	0.041	0.279
20	300	0.131	0.017	100%	0.068	0.210
30	10	0.209	0.050	100%	0.018	0.645
30	60	0.140	0.022	100%	0.030	0.328
30	100	0.136	0.019	100%	0.040	0.279
30	300	0.131	0.011	100%	0.068	0.210

Table 4.5*: Simulation study: Median R^2_h estimates based on 250 simulations for each scenario, where true values set to: R^2_h =0.30 and R^2_{ht} =0.30.*

4.2.1.3 Individual level surrogacy R^2_h: differing strengths of surrogacy

I now discuss the impact of differing strengths of surrogacy at the trial and individual levels on individual level surrogacy estimation.

In Table 4.6 the values of R_h^2 when they were set to 0.64 were similar whether R_{ht}^2=0.90 or R_{ht}^2=0.30. Even though results where surrogacy strengths differ were consistently lower, this bias was always small. This suggests that there was little bias imposed on R_h^2 where surrogacy strengths disagreed. The same direction and degree of bias was also observed where R_h^2=0.30 and trial level was set to be strong, see Appendix A Table A.5.

Number	Trial size	Surrogacy strong both levels			Surrogacy strong R_h^2, weak R_{ht}^2		
		R_h^2 =0.6	IQR R_h^2	% cover CIs	R_h^2 =0.64	IQR R_h^2	% cover CIs
5	10	0.34	0.154	100%	0.336	0.172	100%
5	60	0.30	0.086	100%	0.302	0.078	100%
5	100	0.30	0.063	99%	0.304	0.064	99%
5	300	0.30	0.058	96%	0.301	0.056	97%
10	10	0.33	0.107	100%	0.330	0.106	100%
10	60	0.29	0.063	100%	0.300	0.058	100%
10	100	0.30	0.056	100%	0.294	0.052	100%
10	300	0.29	0.040	99%	0.298	0.046	99%
20	10	0.34	0.072	100%	0.327	0.082	100%
20	60	0.29	0.039	100%	0.292	0.033	100%
20	100	0.29	0.036	100%	0.291	0.039	100%
20	300	0.29	0.031	100%	0.291	0.033	100%
30	10	0.34	0.071	100%	0.327	0.073	100%
30	60	0.29	0.033	100%	0.290	0.031	100%
30	100	0.29	0.027	100%	0.290	0.031	100%
30	300	0.29	0.025	100%	0.288	0.025	100%

Table 4.6: *Simulation study: Median R^2_h estimates based on 250 simulations for each scenario, where true values set to: R^2_h =0.64 and R^2_{ht} =0.90 or 0.30.*

The coverages were comparable across settings with most returning 100% coverage. The IQRs of R_h^2 in the simulations were also consistent in either scenario.

4.2.1.4 Individual level surrogacy R^2_h: non-proportional odds

As can be seen in Table 4.7, when we compare proportional odds against non-proportional odds there was little difference in R_h^2 estimates. Non-proportional odds results were generally lower but within approximately 0.01 of the proportional odds scenario regardless of the size or number of trials.

Number of trials	Trial size	Proportional			Non proportional		
		R_h^2 =0.64	IQR R_h^2	% cover CIs	R_h^2 =0.64	IQR R_h^2	% cover CIs
5	10	0.347	0.154	100%	0.344	0.160	100%
5	60	0.305	0.086	100%	0.310	0.087	99%
5	100	0.300	0.063	99%	0.308	0.063	100%
5	300	0.304	0.058	96%	0.302	0.058	95%
10	10	0.337	0.107	100%	0.322	0.114	100%
10	60	0.293	0.063	100%	0.295	0.051	100%
10	100	0.302	0.056	100%	0.296	0.051	100%
10	300	0.294	0.040	99%	0.288	0.050	100%
20	10	0.342	0.072	100%	0.333	0.078	100%
20	60	0.297	0.039	100%	0.294	0.038	100%
20	100	0.293	0.036	100%	0.294	0.033	100%
20	300	0.292	0.031	100%	0.292	0.031	100%
30	10	0.340	0.071	100%	0.330	0.066	100%
30	60	0.295	0.033	100%	0.292	0.036	100%
30	100	0.294	0.027	100%	0.290	0.030	100%
30	300	0.293	0.025	100%	0.288	0.021	100%

Table 4.7: Simulation study: Median R^2_h estimates based on 250 simulations for each scenario, where true values set to: R^2_h =0.64 and R^2_{ht} =0.90. Comparing results for proportional odds and non-proportional odds.

The coverage and IQRs of the proportional odds and non-proportional odds estimates were comparable between scenarios.

4.2.1.5 Individual level surrogacy R^2_h: ceiling affect

The results of the simulation study presented for R_h^2 show the impact of loss of information that occurs when a binary surrogate and ordinal true outcome represent underlying continua. The values of R_h^2 returned in these settings were just less than half that set in the continuous setting for both R_h^2=0.64 and R_h^2=0.30. I investigated the simulation scenario where R_h^2=1 at the underlying continuum (perfect surrogacy

at the individual level) to see if there was an upper-bound on how useful a binary surrogate can be for an ordinal true outcome.

Number of trials	Trial size	R_h^2	IQR R_h^2	% Cover CIs	lower 95% CI	upper 95% CI
5	10	0.521	0.174	100%	0.109	0.856
5	60	0.531	0.155	93%	0.332	0.714
5	100	0.516	0.156	86%	0.368	0.648
5	300	0.490	0.155	53%	0.413	0.566
10	10	0.500	0.113	100%	0.108	0.840
10	60	0.502	0.098	99%	0.315	0.672
10	100	0.504	0.097	93%	0.358	0.634
10	300	0.488	0.106	70%	0.410	0.566
20	10	0.493	0.086	100%	0.107	0.836
20	60	0.501	0.068	100%	0.312	0.669
20	100	0.491	0.066	98%	0.351	0.624
20	300	0.489	0.066	88%	0.409	0.568
30	10	0.484	0.071	100%	0.105	0.823
30	60	0.493	0.055	100%	0.310	0.664
30	100	0.491	0.056	100%	0.351	0.621
30	300	0.479	0.051	96%	0.401	0.554

Table 4.8: Simulation study: Median R^2_h estimates based on 250 simulations for each scenario, where true values set to: R^2_h =1 and R^2_{ht} =0.90..

The estimates for R_h^2=1 seemed to converge on approximately 0.48 as trial sizes increased, and to a lesser extent as the number of trials increased, see Table 4.8. This suggested that the most useful a binary surrogate for an ordinal true outcome can be, at the individual level, had a ceiling of around 0.48.

As the number of patients increase the coverage is increasingly poor, except where there are a large number of trials. The IQRs are much larger in this setting than in previous settings suggesting that the R_h^2 values between simulations vary a lot. Also, the confidence intervals decrease in size as the number of patients increase (see the median bounds). Where the IQRs are large and the intervals decrease it is much more likely that the interval will not contain the large sample approximation of R_h^2 which

explains why the coverage decreases. Overall in the case of 'perfect' surrogacy the intervals do not appear to work well, however, this is a very particular scenario that is unlikely to occur in practice. Other settings do not have such large IQRs and are unlikely to suffer the same issue to the same degree.

4.2.1.6 Individual level surrogacy R^2_h: comparison to other methodology

No publications on information theory surrogacy evaluation for ordinal surrogates were available in the literature. However, it was possible to compare my results to the published findings in the binary-binary and binary-continuous settings.

4.2.1.6.1 Comparison to the binary-binary setting

In the binary-binary setting, (Tilahun et al., 2008a), results converge to approximately 0.21 compared to 0.29 in the binary-ordinal case when R^2_h=0.64 and R^2_{ht}=0.90 at the underlying continua. This suggest that an ordinal true outcome is more desirable when investigating surrogacy than with a binary true outcome. This is logical given the additional loss of information in the binary-binary setting.

However, there is another possible explanation for the difference. The information theory results in the binary-binary setting were produced using the LRF based on one full model for all trials as opposed to the summation of the LRF for separate models for each trial (see section 3.4.3). Tilahun et al., 2008a used a full model method according to the freely available software released with this publication (I-Biostat, 2015). Therefore, only one statistic of the amount of information gained about the true outcome (one LRF calculation) after adjustment for the surrogate was used rather than a summation of this statistic across individual trials. Therefore, they relied heavily on the assumption that association between the surrogate and true outcome was constant over trials. This could also be an explanation for the differences seen in R^2_h results between the binary-binary setting and my results for the binary-ordinal setting.

4.2.1.6.2 Comparison to the binary-continuous setting

In the binary-continuous setting values of R^2_h converged to approximately 0.16, Pryseley et al. (2007), compared to 0.29 in the binary-ordinal setting when R^2_h=0.64

and R^2_{ht}=0.90 at the underlying continua. This was again much lower than in the binary-ordinal simulation results. This is contrary to expectation since the binary-continuous setting suffers less loss of information than the binary-ordinal setting.

4.2.1.6.3 Comparison conclusions

Using these comparisons, I could suggest that the results of the binary-ordinal setting showed that a binary surrogate for an ordinal true outcome was more informative than other discrete outcome settings. However, it was equally likely that the use of a multi-trial method for individual level surrogacy in this case was the reason for the improvements in these results. Suggesting that a multi-trial method is superior to the methods used in the binary-binary and binary-continuous publications.

4.2.1.7 Individual level surrogacy R^2_h: Conclusions

The results of this simulation study suggest that the loss of information imposed by dichotomising a continuous surrogate and categorising an ordinal true outcome was substantial. A binary surrogate for an ordinal true outcome appears to be half as informative as their continuous counterparts. It would be interesting to see if any improvement on this ceiling would be possible with an ordinal surrogate as opposed to a binary one, as this would provide more information (this will be discussed in 6.2.1.6). However, it should be noted that an ordinal surrogate is not always better than a binary surrogate if the binary representation more naturally describes the status of the measure it is representing (Burzykowski et al., 2005).

Median R^2_h estimates were fairly consistent regardless of the size or number of trials. Although, an increase in trial size slightly improved estimation of R^2_h. Differing strengths of surrogacy at the trial and individual level and the presence of non-proportional odds had little impact on R^2_h estimates. Research has shown that the proportional odds model is capable of handling divergences from the proportional odds assumption well (McHugh et al., 2010a). This may explain why individual level surrogacy assessment performed well regardless of the presence of non-proportional odds.

The confidence intervals had 100% coverage in nearly all scenarios, that are likely to occur in practice, suggesting that these are conservative. The median 95% bounds

suggested that the intervals do not cover the whole parameter space and are relatively narrow in some scenarios suggesting that they are useful to some degree.

In the case of "perfect" surrogacy, see section 4.2.1.5, the intervals do not work well – however this is a very particular case that is unlikely to occur in practice and other scenarios do not suffer the same issues to the same degree.

The coverage at the individual level was based on the large scale approximation, rather than the true value of R_h^2, therefore, these were not ideally formatted, see 4.1.3.3.3. However, the overall impression from the coverage results is that these intervals are too conservative in the general case and alternatives should be considered.

4.2.2 Results: trial level surrogacy

In this section I describe:

- the scenario where surrogacy was strong at both the trial and individual levels;
- where surrogacy was weak at both levels;
- where the strength of surrogacy disagrees at the trial and individual levels;
- where non-proportional odds were present;
- and finally, compare the penalised likelihood to the trial removal technique for dealing with separation.

4.2.2.1 Trial level surrogacy R^2_{ht}: strong surrogacy

Consider Table 4.9, the results for trial level surrogacy when surrogacy was set to be strong at both levels (R_{ht}^2 simulated to be 0.90) are presented.

Table 4.9 shows that there was overestimation where there were five trials and moderate to large trial sizes but as the number of trials increased the median R_{ht}^2 displayed greater underestimation. This was worse for smaller trial sizes. On the whole, the coverage of the confidence intervals improved as the number and size of trials increased – with approximately 95% coverage for larger numbers of patients per trial. However, in the case of small numbers of patients per trial and larger trial

numbers the coverage is poor, this is likely because of the relatively large amount of bias in $\boldsymbol{R^2_{ht}}$ results in these scenarios rather than issues with the intervals themselves.

Number of trials	Trial size	$\boldsymbol{R^2_{ht}}$	IQR $\boldsymbol{R^2_{ht}}$	% Cover CIs	Median CI lower	Median CI upper
5	10	0.781	0.367	82%	0.120	0.988
5	60	0.923	0.146	93%	0.408	0.998
5	100	0.924	0.135	92%	0.411	0.998
5	300	0.949	0.111	93%	0.513	0.999
10	10	0.616	0.304	68%	0.121	0.922
10	60	0.838	0.153	84%	0.412	0.979
10	100	0.862	0.132	91%	0.462	0.984
10	300	0.900	0.090	95%	0.553	0.990
20	10	0.571	0.203	45%	0.207	0.843
20	60	0.803	0.120	76%	0.503	0.947
20	100	0.831	0.098	86%	0.552	0.957
20	300	0.870	0.082	95%	0.625	0.969
30	10	0.549	0.169	21%	0.249	0.791
30	60	0.798	0.100	69%	0.560	0.928
30	100	0.829	0.092	83%	0.611	0.942
30	300	0.865	0.065	94%	0.672	0.957

Table 4.9: *Simulation study: Median R^2_{ht} estimates based on 250 simulations for each scenario, in the binary-ordinal setting where true values set to: R^2_h =0.64 and R^2_{ht} =0.90.*

The size of the IQRs of R^2_{ht} were larger than that experienced at the individual level in the same scenario (see section 4.2.1.1). Given that the two stage trial level surrogacy approach was used, second stage models were based on a sample of size equal to the number of trials. Conversely, a one stage approach (as used at the individual level) would have been based on the number of trials multiplied by the size of the trials. Therefore, the larger ranges in the distributions is unsurprising given that trial level surrogacy was based on two stages. For instance, in the scenario of 30 trials and 300 patients the IQR in R^2 results at the individual level was 0.025as opposed to 0.065 at the trial level.

4.2.2.1.1 Comparison to continuous-continuous setting: strong surrogacy

Theoretically, at the trial level, the results for the binary-ordinal setting and the setting where a continuous surrogate and continuous true outcome (continuous-continuous setting) were modelled, should be the same. We can compare binary-ordinal results to those for the continuous-continuous setting. In the continuous setting there was also overestimation when there were only five trials, see Table 4.10. However, the underestimation seen in the binary-ordinal setting for ten or above trials was not present in the continuous-continuous setting. Where there were 30 trials and 300 patients the median R_{ht}^2 was 0.865 in the binary-ordinal setting and 0.902 in the continuous setting. Large underestimation can also be seen for smaller trial sizes and number of trials in the binary-ordinal setting. This underestimation was not as pronounced in the continuous-continuous setting.

		Binary-ordinal		Continuous-continuous	
Number of trials	Trial size	R^2_{ht}	IQR R^2_{ht}	R^2_{ht}	IQR R^2_{ht}
5	10	0.781	0.367	0.931	0.165
5	60	0.923	0.146	0.950	0.103
5	100	0.924	0.135	0.953	0.095
5	300	0.949	0.111	0.959	0.069
10	10	0.616	0.304	0.840	0.137
10	60	0.838	0.153	0.909	0.096
10	100	0.862	0.132	0.916	0.072
10	300	0.900	0.090	0.919	0.068
20	10	0.571	0.203	0.835	0.100
20	60	0.803	0.120	0.900	0.065
20	100	0.831	0.098	0.902	0.054
20	300	0.870	0.082	0.910	0.058
30	10	0.549	0.169	0.826	0.078
30	60	0.798	0.100	0.895	0.053
30	100	0.829	0.092	0.900	0.050
30	300	0.865	0.065	0.902	0.054

Table 4.10: Simulation study: Median R^2_{ht} estimates based on 250 simulations for each scenario, in binary-ordinal and continuous-continuous setting where true values set to: R^2_h =0.64 and R^2_{ht} =0.90.

4.2.2.1.2 Underestimation R^2_{ht}: strong surrogacy

Given that the continuous setting does not suffer such large issues of bias it seemed that this was at least partially driven by the discrete nature of the outcomes in the binary-ordinal setting.

The two stage nature of the information theory approach at the trial level and categorisation of outcomes led to inefficiencies in estimation. Recall that at the first stage of modelling R^2_{ht} treatment effects were estimated for each trial for both the surrogate and the true outcome. These estimates were then regressed in the models at the second stage to calculate the R^2_{ht}. This two stage approach was introduced as an alternative to the computationally burdensome full joint mixed model of the meta-

analytical approach (see section 2.2.1). The two stage approach was effectively: taking a large computational problem; partitioning it into smaller components; solving for each component; and combining the results of these to give an overall answer.

Molenberghs et al. (2011) investigated the impact of a similar method on statistical efficiency and found that when the number of partitions was large compared to the size of the partitions, inefficiency occurs. Partitions in this case were trials. In Table 4.10 as the number of trials increased and sizes decreased the bias worsened which was in keeping with the results of Molenberghs et al. (2011).

On top of the inefficiency due to the two stage approach, binary or ordinal outcomes are less efficient at providing parameter estimation than continuous outcomes. In our case the response variables of the stage one models have been dichotomised or categorised to create binary and ordinal response outcomes. (In either case the discrete variables were regressed on treatment to estimate treatment effects for each trial.) Taylor et al. (2006) and Taylor and Yu (2002) showed that where regression variables were categorisations of continuous variables inefficiency occurs in estimation. This inefficiency is much worse for binary outcomes as opposed to categorical outcomes. This result suggested that the dichotomisation of the binary surrogate contributed more to the inefficient results than the categorisation of the ordinal true outcome.

In summary, the bias in results was due to inefficiency imposed by a two stage approach compounded by the use of binary and ordinal outcomes. This was substantiated by the larger IQRs in the R^2_{ht} results than for R^2_h (where a one stage approach was used). This result was also true in the continuous-continuous setting, data not shown. Furthermore, the IQRs of R^2_{ht} in the continuous-continuous setting were narrower than in the binary-ordinal setting, see Table 4.10.

Consider Table 4.11, I present additional settings where each trial has tenfold more patients (3000 patients) than the highest scenario currently considered. The median R^2_{ht} results were much closer to the true strength of surrogacy in these scenarios and the widths of the IQRs were closer to that in the continuous setting. Therefore, the

inefficient binary-ordinal setting requires larger samples per partition (trial) to gain unbiased estimates.

		binary-ordinal	
Number of trials	Trial size	R^2_{ht}	IQR R^2_{ht}
5	3000	0.959	0.079
10	3000	0.902	0.078
20	3000	0.896	0.062
30	3000	0.890	0.057

Table 4.11: Additional simulation scenarios: Median R^2_{ht} estimates based on 250 simulations for each scenario, where R^2_h =0.64 and R^2_{ht}=0.90 and trial size was 3000.

4.2.2.1.3 Overestimation R^2_{ht}: strong surrogacy

There was some evidence of overestimation in Table 4.10 for small numbers of trials. This issue was referenced in row three of Table 4.3 and the reason for this bias will be full explained in section 4.2.2.2.5 in reference to weak surrogacy strengths where this issue is more pronounced. However, the overestimation occurs in the case of strong surrogacy for small numbers of trials for the same reason.

4.2.2.2 Trial level surrogacy R^2_{ht}: weak surrogacy

Here I discuss the case where surrogacy is weak at both levels, $R^2_h = R^2_{ht}$=0.30, I expect the trial level results to be in the region of 0.30.

The results of this section were more complex than others therefore, I will: first, describe the results; compare them to the underlying continuous-continuous setting; discuss them in relation to the two conflicting biases; and then conclude.

4.2.2.2.1 Results R^2_{ht}: weak surrogacy

Consider Table 4.12, when surrogacy was weak there was extremely large overestimation where trial numbers were small which worsened as the size increased. As the number of trials increased estimation improved where trial sizes were large but then displayed underestimation where trial sizes were small.

Number of trials	Trial size	R^2_{ht}	IQR R^2_{ht}	% cover CIs	Median CI lower	Median CI upper
5	10	0.561	0.539	74%	0.011	0.959
5	60	0.596	0.485	76%	0.016	0.965
5	100	0.670	0.491	71%	0.038	0.976
5	300	0.658	0.473	79%	0.033	0.974
10	10	0.327	0.326	76%	0.005	0.790
10	60	0.440	0.366	91%	0.023	0.852
10	100	0.431	0.360	95%	0.020	0.847
10	300	0.401	0.385	95%	0.014	0.832
20	10	0.238	0.227	92%	0.009	0.602
20	60	0.324	0.299	95%	0.035	0.678
20	100	0.329	0.202	97%	0.037	0.682
20	300	0.333	0.224	96%	0.039	0.685
30	10	0.211	0.171	97%	0.017	0.510
30	60	0.283	0.176	88%	0.047	0.583
30	100	0.303	0.154	92%	0.057	0.601
30	300	0.313	0.164	94%	0.063	0.610

Table 4.12: Simulation study: median R^2_{ht} estimates based on 250 simulations for each scenario, in the binary-ordinal setting where true values set to: R^2_h =0.30 and R^2_{ht}=0.30.

The size of the IQRs of R^2_{ht} were large compared to where surrogacy was set to be strong, see Table 4.9. The IQRs decreased as the number of trials increased but remains much larger in relation to strong surrogacy strength scenarios. For instance, in the information-rich scenario of 30 trials and 300 patients the IQR when surrogacy was strong was 0.065 as opposed to 0.164 where surrogacy was set to be weak. The coverage improved as the number of trials increased, but was quite poor for small numbers of trials, this was again likely due to the increased bias in the R^2_{ht} results.

4.2.2.2.2 Comparison to continuous-continuous setting: weak surrogacy

In order to understand these results it was necessary again to compare them to those for the continuous-continuous setting.

In Table 4.13 the pattern of overestimation seen in the binary-ordinal setting was consistent with the continuous-continuous setting.

		Binary-ordinal setting		Continuous-continuous setting	
Number of trials	Trial size	R^2_{ht}	IQR R^2_{ht}	R^2_{ht}	IQR R^2_{ht}
5	10	0.561	0.539	0.600	0.453
5	60	0.596	0.485	0.659	0.399
5	100	0.670	0.491	0.676	0.539
5	300	0.658	0.473	0.683	0.440
10	10	0.327	0.326	0.438	0.298
10	60	0.440	0.366	0.446	0.347
10	100	0.431	0.360	0.402	0.376
10	300	0.401	0.385	0.429	0.324
20	10	0.238	0.227	0.349	0.212
20	60	0.324	0.299	0.353	0.199
20	100	0.329	0.202	0.327	0.213
20	300	0.333	0.224	0.355	0.222
30	10	0.211	0.171	0.326	0.172
30	60	0.283	0.176	0.329	0.179
30	100	0.303	0.154	0.358	0.210
30	300	0.313	0.164	0.336	0.204

Table 4.13: Median R^2_{ht} estimates based on 250 simulations for each scenario, in continuous-continuous setting where true values set to: R^2_h =0.30 and R^2_{ht} =0.30.

In both settings there was large overestimation as the number of trials decreased. The median R^2_{ht} result was as high as 0.658 for five trials and 300 patients against a true value of 0.30. A result of 0.658 would represent the case of a moderately good surrogate, which could lead to a false conclusion of a valid surrogate. The continuous-continuous setting showed overestimation for every scenario.

4.2.2.2.3 Comment R^2_{ht}: weak surrogacy

Before I go on to discuss the reason for this overestimation, note that compared to the continuous-continuous setting the binary-ordinal results were again showing

lower results as the number of trials increased. This was still present in larger trial sizes but was a more serious issue for smaller trial sizes. Therefore, the binary-ordinal setting results seemed to be more accurate only because they were suffering two opposing forms of bias (see rows two and three of Table 4.3).

4.2.2.2.4 Underestimation R^2_{ht}: weak surrogacy

The reason for the underestimation in the binary-ordinal setting was again due to the inefficiency imposed by the use of a binary surrogate and ordinal true outcome instead of continuous outcomes, as described in section 4.2.2.1.2. Again, if we add an extra scenario to our simulation where the largest number of patients was increased tenfold (3000 patients) we find that the results in the binary-ordinal setting more closely resemble those for the larger trial sizes of the continuous-continuous setting (see Table 4.14). Given that a larger sample size returns more comparable results to the continuous setting, it was fair to conclude that this underestimation was caused by the inefficiency of the binary and ordinal outcomes.

		Binary-ordinal	
Number of trials	Trial size	R^2_{ht} =0.30	IQR R^2_{ht}
5	3000	0.591	0.428
10	3000	0.491	0.300
20	3000	0.360	0.245
30	3000	0.326	0.191

Table 4.14: Additional simulation scenarios: Median R^2_{ht} estimates based on 250 simulations for each scenario, where $R^2_h = R^2_{ht} = 0.30$ and trial size was 3000

4.2.2.2.5 Overestimation R^2_{ht}: weak surrogacy

I now discuss the reasons for overestimation in the binary-ordinal and continuous-continuous settings.

The information theory approach at the trial level was effectively based on the coefficient of determination (R^2) of the second stage models. In this case, the coefficient of determination is a measure of the ability of the surrogate treatment

effect estimates to predict those for the true outcome. Better predictive ability corresponds to a higher estimate of the strength of surrogacy.

We can see graphically in Figure 4.2 an artificial example assembled using the simulation code. First, let us consider the case where there are 30 trials. Here, where surrogacy is strong ($R^2_{ht} = 0.90$) treatment effect estimates are very close to the regression line. Conversely, the residuals are larger for the case of lower surrogacy and 30 trials, and the points are more widely scattered, i.e. have poor predictive ability. If we select only five of these points, to represent the situation where there were five trials, for strong surrogacy the regression is still representative of that for a larger number of trials, i.e. the residuals are small.

The surrogacy estimates are $\widehat{R^2_{ht}}$= 0.757 for both five and 30 trials, where we are expecting results of 0.90.

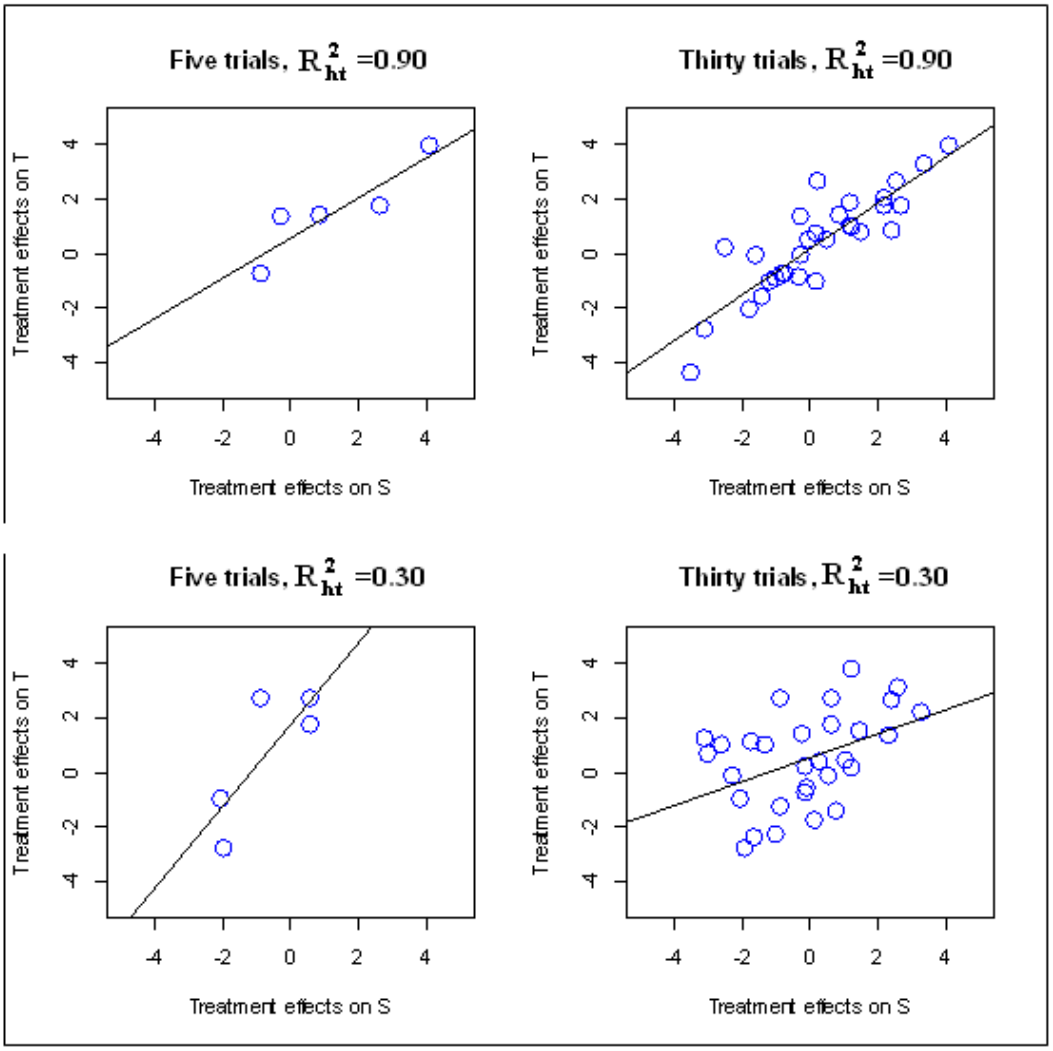


Figure 4.2: Regressions of treatment effect estimates on the true outcome regressed on the surrogate for: five/thirty trials and high surrogacy (top left and right resp.); five/thirty trials and weak surrogacy (bottom left and right respectively)

However, when surrogacy was weak, since there was much more variability in the treatment effect estimates for the five selected trials, these do not provide good estimation. The regression based on the five selected trials was not representative of that obtained for thirty trials. In these models the true relationship is not fitted and the model instead fits too closely to these inadequate data (in this case because there were too few data points). Hence, there were smaller residuals than in the 30 trials case and the model appeared to have higher predictive ability/surrogacy. This was simply due to the fact that there were fewer points and is an example of overfitting.

Where there were 30 points (30 trials), overfitting does not occur and estimates closer to the true value of surrogacy were obtained. Here, $\widehat{R_{ht}^2}$=0.638 for five trials and $\widehat{R_{ht}^2}$=0.156 for thirty trials where we were expecting results of 0.30. This example highlights why it is difficult to estimate surrogacy effectively when there are lower strengths of surrogacy and small numbers of trials.

4.2.2.2.6 Conclusion R^2_{ht}: weak surrogacy

Due to the nature of the information theory, when there is weak surrogacy, overfitting for small numbers of trials leads to inflated estimation. This explains the overestimation witnessed in results for small numbers of trials, the much wider IQRs in R_{ht}^2 and worse coverage of confidence intervals. The results in Table 4.13 show that overestimation was present even in the case of 30 trials.

I have also found that, in keeping with the results of strong surrogacy, the binary-ordinal setting displays underestimation compared to the continuous-continuous setting due to the inefficiency of discrete variables as opposed to continuous variables.

4.2.2.3 Trial level surrogacy R^2_{ht}: differing strengths of surrogacy

In Table 4.15 we can compare the results where surrogacy was strong at both levels to the case where surrogacy was strong at the trial level but weak at the individual level. Generally, having different strengths of surrogacy at both levels does not affect estimation of R_{ht}^2. However, as the number of trials increased, if trial sizes were small, there was greater underestimation for disagreement in surrogacy strengths.

In the converse case, where surrogacy was weak at the trial level and strong at the individual level, we also see an impact on the results where surrogacy disagrees. Again, this was worse for larger trial numbers and small trial sizes, see Appendix A Table A.4.

Number of trials	Trial size	Surrogacy strong both levels			Surrogacy strong R^2_{ht}, weak R^2_h		
		R^2_{ht} =0.90	IQR R^2_{ht}	% cover CIs	R^2_{ht} =0.90	IQR R^2_{ht}	% cover CIs
5	10	0.781	0.367	82%	0.753	0.325	85%
5	60	0.923	0.146	93%	0.902	0.174	88%
5	100	0.924	0.135	92%	0.913	0.154	93%
5	300	0.949	0.111	93%	0.947	0.101	92%
10	10	0.616	0.304	68%	0.531	0.320	74%
10	60	0.838	0.153	84%	0.826	0.159	88%
10	100	0.862	0.132	91%	0.852	0.148	85%
10	300	0.900	0.090	95%	0.896	0.085	96%
20	10	0.571	0.203	45%	0.451	0.226	38%
20	60	0.803	0.120	76%	0.774	0.136	71%
20	100	0.831	0.098	86%	0.817	0.104	83%
20	300	0.870	0.082	95%	0.861	0.081	94%
30	10	0.549	0.169	21%	0.447	0.194	17%
30	60	0.798	0.100	69%	0.757	0.103	55%
30	100	0.829	0.092	83%	0.798	0.088	79%
30	300	0.865	0.065	94%	0.857	0.066	92%

Table 4.15: Simulation study: Median R^2_{ht} estimates based on 250 simulations for each scenario, where true values set to: R^2_{ht} =0.90 and R^2_h =0.64 or 0.30.

4.2.2.4 **Trial level surrogacy R^2_{ht}: non-proportional odds**

I will now discuss the results where the odds of the ordinal true outcome were not proportional. The binary-ordinal results for the non-proportional odds scenario were similar to the proportional odds scenario, see Table 4.16.

This suggests that the presence of deviations from the proportional odds assumption on an ordinal true outcome has little impact on the estimation of surrogacy at the trial level.

Number of trials	Trial size	Proportional			Non proportional		
		R^2_{ht} =0.90	IQR R^2_{ht}	% cover CIs	R^2_{ht} =0.90	IQR R^2_{ht}	% cover CIs
5	10	0.781	0.367	82%	0.791	0.311	86%
5	60	0.923	0.146	93%	0.916	0.152	88%
5	100	0.924	0.135	92%	0.920	0.124	91%
5	300	0.949	0.111	93%	0.945	0.114	93%
10	10	0.616	0.304	68%	0.595	0.304	72%
10	60	0.838	0.153	84%	0.837	0.133	86%
10	100	0.862	0.132	91%	0.858	0.129	89%
10	300	0.900	0.090	95%	0.881	0.123	93%
20	10	0.571	0.203	45%	0.533	0.221	41%
20	60	0.803	0.120	76%	0.808	0.120	79%
20	100	0.831	0.098	86%	0.827	0.101	79%
20	300	0.870	0.082	95%	0.858	0.080	94%
30	10	0.549	0.169	21%	0.525	0.164	21%
30	60	0.798	0.100	69%	0.788	0.098	64%
30	100	0.829	0.092	83%	0.833	0.077	81%
30	300	0.865	0.065	94%	0.858	0.080	94%

Table 4.16*: Simulation study: Median R^2_{ht} estimates based on 250 simulations for each scenario, where true values set to: R^2_{ht} =0.90 and R^2_h =0.64. Comparing results for proportional odds and non-proportional odds.*

4.2.2.5 Trial level surrogacy R^2_{ht}: dealing with separation

Recall that we discussed issues of separation in section 3.6 and how this impacts surrogacy estimation. We also discussed how this was dealt with in the simulation in section 4.1.3.3. For an overview of this issue see row four of Table 4.3.

4.2.2.5.1 Occurrence of separation: trial level

The separation for binary outcomes in the simulation ranged from 2.5% to 79.8%. The average was 23.4%. Separation in the ordinal case was recorded independently for complete and quasi-complete separation. The amount of ordinal complete separation ranged from 0% to 28.6% with an average of 4.5%. Quasi-complete separation for ordinal outcomes ranged from 0% to 23.3% with an average of 3.2%.

Hence, under this random simulation process, separation occurred frequently and needed to be addressed.

4.2.2.5.2 Comparison of penalized likelihood and trial removal technique: separation at the trial level

Two techniques were taken to deal with the issues of separation. The first was to apply a penalized likelihood technique which allows trials with separation to be retained in analysis. The second technique was to remove trials where separation occurs which led to loss of information.

In Table 4.17 the penalized likelihood technique was compared to the removal of trials technique. Focusing, firstly, on the case of 20 and 30 trials the estimates of surrogacy for smaller trial sizes were much better using the penalized likelihood technique. In the trial removal technique these estimates were extremely poor where there were only ten patients per trial.

The results for lower numbers of trials (five or ten trial scenarios) also send the same message, although this is less obvious at first. When we consider the median R_{ht}^2, the results appear to be better or similar for the trial removal technique where trial sizes were small. However, in the case of as few as ten patients per trial, it was likely that separation occurred very frequently. In the case of fewer than three trials being available for the analysis for the removal trial technique the simulation was set to return a null value (see section 4.1.3.2). The simulation was run until 250 results were obtained for both the penalized likelihood and removal technique i.e. 250 datasets were simulated where three or more trials were available.

In the five trials of ten patients scenario, 90.3% of datasets had less than three trials available for analysis and therefore null results were returned for the trial removal technique. In the ten trials and ten patients per trial scenario 46.1% of simulated datasets returned a null value, see Table 4.17. Where simulations were not rejected, for having only three or fewer trials, generally a lot fewer than the stated number of trials were available for analysis. I found that as the number of trials removed increased so too did the value of R_{ht}^2 using the trial removal technique. This was due to the overestimation outlined in row three of Table 4.3 and discussed in section

4.2.2.2.5. Therefore, where there were small numbers of trials R^2_{ht} estimates for the trial removal technique were artificially high due to bias and were not a result of more effective estimation compared to the penalized likelihood technique.

Number of trials	Trial size	Penalized likelihood technique		Trial removal technique			
		R^2_{ht} =0.90	IQR R^2_{ht}	R^2_{ht} =0.90	IQR R^2_{ht}	% failures	Median trial No.
5	10	0.781	0.367	0.803	0.464	90.3%	4
5	60	0.923	0.147	0.930	0.159	14.4%	5
5	100	0.924	0.135	0.934	0.136	7.1%	5
5	300	0.949	0.111	0.948	0.114	0.8%	5
10	10	0.616	0.304	0.633	0.546	46.1%	5
10	60	0.838	0.153	0.833	0.263	0.0%	9
10	100	0.862	0.132	0.847	0.161	0.0%	9
10	300	0.900	0.090	0.895	0.101	0.0%	10
20	10	0.571	0.203	0.411	0.420	2.7%	7
20	60	0.803	0.120	0.793	0.191	0.0%	17
20	100	0.831	0.098	0.826	0.146	0.0%	18
20	300	0.870	0.082	0.871	0.080	0.0%	19
30	10	0.549	0.169	0.278	0.383	0.4%	11
30	60	0.798	0.100	0.783	0.162	0.0%	25
30	100	0.829	0.092	0.823	0.103	0.0%	27
30	300	0.865	0.065	0.866	0.066	0.0%	29

Table 4.17: *Simulation study: Median R^2_{ht} estimates based on 250 simulations for each scenario, where true values set to: R^2_{ht} =0.90 and R^2_h =0.64. Comparing penalized likelihood technique against trial removal technique (trial removal technique results include the % of time the calculation of R^2_{ht} was not possible and the median number of trials available for analysis when it was).*

This finding was supported by the fact that the IQRs for the penalised likelihood technique were narrower than the trial removal technique.

As the number of patients increased the benefits of the penalized likelihood technique were less evident in the median R^2_{ht} results. However, for these scenarios the penalized technique had narrower IQRs of R^2_{ht}. Meaning that in general results for the penalized likelihood technique were more precise. This was because the penalized likelihood technique was based on the full dataset in each case, whereas the trial removal technique will often be based on a much reduced dataset due to the removal of trials data where separation occurs. Often fewer than half of the trials were retained in analysis (see Table 4.17). This represents a huge loss of information when the trial removal technique was implemented. Even where there were large numbers of patients per trial the median number of trials retained in the removal technique was generally lower than that simulated, see Table 4.17.

In comparison the penalized likelihood allows the retention of all trials in the analysis and does not suffer loss of information. The only time trials are removed from analysis for the penalized likelihood approach is when failure in the pordlogist model occurs for the ordinal outcome, however this only happened 0.5% of the time.

Hence, I concluded that the penalized likelihood technique was far superior to a trial removal technique as it provides better estimation and does not suffer loss of information. This valuable methodological advancement has solved a problematic issue in discrete information theory surrogacy evaluation.

4.2.2.6 Trial level surrogacy R^2_{ht}: comparison to other methodology

4.2.2.6.1 Weak surrogacy: comparison

I have highlighted the issue of overestimation in the information theory approach where there were small numbers of trials which mostly impacts lower strengths of surrogacy. Only one previous publication assesses lower strengths of surrogacy Burzykowski et al.(2005). This publication used the meta-analytical as opposed to the information theory approach but the results should be comparable to ours since these two approaches are very similar. In Burzykowski et al.(2005) surrogacy was set to be 0.50 at the trial level for continuous outcomes. The summary R^2 estimates were in the region of 0.54 when results of 0.50 were expected. This simulation was based on 25 trials with 50 plus patients in each trial. This was a similar amount of

overestimation to that which was observed in my study for the continuous-continuous setting, see Table 4.10 for 30 trials. Therefore, these results go some way to corroborating my findings.

4.2.2.6.2 Discrete outcome information theory approaches

There have been two previous papers on information theory surrogacy evaluation for discrete outcomes. One paper discussed the continuous-binary setting, (Pryseley et al., 2007) and one the binary-binary setting, (Tilahun et al., 2008b). Both settings showed much worse underestimation than in the binary-ordinal setting when the true value was 0.90 (see Table 4.3). For 30 trials and 300 patients and strong surrogacy, the median R^2_{ht} value was 0.75 in the binary-continuous setting and 0.764 in the binary-binary setting, as opposed to 0.865 in my binary-ordinal setting.

I believe this was because these authors have not taken account of the issue of separation. The code available for these publications on the authors website gives no indication of adjustment for or removal of separation occurrences (I-Biostat, 2015). Furthermore, when problematic separation trials were not removed from my analysis the bias in results was similar to that reported in these publications. Hence, I attribute their observed bias in the binary-binary and binary-continuous settings to the unresolved problem of separation. The serious bias in these publications demonstrates the advantage of the development of the penalized likelihood technique to deal with issues of separation.

4.2.2.7 Trial level surrogacy R^2_{ht}: conclusions

In conclusion, trial level surrogacy though not theoretically affected by the categorisation of continuous outcomes was impacted practically speaking due to inefficiencies in estimation imposed through the use of a two stage approach. This led to underestimation as the number of trials increased and trial sizes decreased. It also led to wider IQRs.

When the number of trials was small the regressions were based on only a few data points (treatment effect estimates for each trial). This led to overfitting in stage two models and the surrogate falsely appeared to have a strong predictive/surrogate potential. Ten or more trial scenarios led to less inflated estimation.

When there was no bias in R^2_{ht} results the confidence intervals showed good coverage, especially for larger trial sizes and numbers of patients per trial.

I have also demonstrated the huge advantage of using a penalized likelihood technique to deal with the occurrence of separation in discrete outcomes for surrogacy evaluation; as opposed to a trial removal technique or ignoring the issue altogether. Separation was found to occur frequently under simulation and to greatly bias the results in previous publications (Pryseley et al., 2007) and (Tilahun et al., 2008b).

Results for differing surrogacy strengths at the trial and individual levels showed that there was some bias in results for small trial sizes which worsened as the number of trials increased. There was little effect on results where non-proportional odds were present.

4.2.3 Results: conclusions

I have shown that the loss of information that impacted R^2_h in the observed binary-ordinal setting was substantial. Where a binary surrogate and an ordinal true outcome were present the estimates of R^2_h in the observed setting were just less than half as informative as those set at the underlying continua. The presence of discordant strength of surrogacy at the trial and individual levels and non-proportional odds have little impact on the estimation of R^2_h.

Considering previous publications for discrete outcomes, it appeared that binary surrogates for ordinal true outcomes are more informative than other discrete scenarios. The alternative explanation was that using a method based on the amount of information gained for each trial separately is superior to methods that calculate the amount of information gained for all trials as a whole.

The confidence intervals at the individual level are conservative, perhaps bootstrap intervals would be more appropriate.

Three issues were identified at the trial level: underestimation; overestimation and separation. These can be considered as issues relating to the two stage nature of the information theory approach. Underestimation and separation were due to or resulted

in difficulties in estimation in stage one models, the former because of inefficiencies and the latter due to modelling issues. Overestimation was due to overfitting in stage two models.

Simulation demonstrated a high frequency of separation and showed the need for a solution to this problem in surrogate evaluation methodology. The use of a penalized likelihood technique resolved the issue of separation, this has not previously been considered in the surrogate literature. The simulation and comparisons to previous published research have shown that techniques that ignore or remove trials that suffer separation were greatly inferior to the penalized likelihood technique. The penalized likelihood technique returns sensible results without suffering loss of information.

4.3 Simulation study: conclusions

I set up a simulation study to investigate the ability of the information theory approach in the binary-ordinal setting to estimate surrogacy. I have carefully investigated the literature in considering the set-up of the simulation study then effectively demonstrated the worth of the information theory methodology in various scenarios.

This simulation study has identified issues and interesting aspects of information theory that have been previously unknown. Where possible, methodological resolutions to this issues have been demonstrated. Overall, extending the information theory approach to the binary-ordinal setting has proved to be a useful addition to the collection of settings already developed.

Chapter 5. Extension of information theory to the case of an ordinal surrogate and binary true outcome: Methodology

In this Chapter I outline how the information theory approach can be extended to the case of an ordinal surrogate and binary true outcome. I: re-introduce the information theory approach for continuous outcomes; show how this can be extended to the ordinal-binary setting at the individual and trial level; and briefly discuss the application of the penalized likelihood method in the ordinal-binary setting.

As in previous chapters I represent a putative surrogate as S, treatment as Z and the true outcome as T. In the multi trial context there are i=1,2,…,N trials, and j=1,2,….,n_i patients per trial. $N_T = \sum_i n_i$, the total number of patients in all trials. Here the ordinal surrogate has V ordered categories.

5.1 The information theory approach: a re-introduction

Information theory involves the study of the amount of information or uncertainty in a random variable. Entropy is a key concept and informs on the amount of uncertainty in a draw from a random variable, i.e. if one outcome is much more likely then we are fairly certain what the result of the draw will be. Another useful concept in information theory is that of the mutual information, which calculates the amount of uncertainty in one variable that can be accounted for by another, see section 3.2 for more details.

Alonso and Molenberghs (2007) proposed using these concepts among others to calculate the amount of uncertainty in the true outcome (treatment effects on T) accounted for by the surrogate (treatment effects on S) at the individual (trial) level. This was an intuitive proposal since surrogacy investigations require knowledge of the amount of information on the true outcome explained by the surrogate. The measures proposed were R_h^2 for the individual level and R_{ht}^2 at the trial level, see

section 3.2. Alonso and Molenberghs (2007) suggested the use of the likelihood reduction factor (LRF) to estimate R_h^2 and R_{ht}^2 in the multi trial setting as it provides a consistent interpretation across settings and ranges in the unit interval. An R^2=0 represents a surrogate of no value and R^2=1 the perfect surrogate.

In the following sections I re-introduce the LRF and show how it can be used to calculate R_h^2 and R_{ht}^2. I then show how the LRF can be applied to the case of an ordinal surrogate and binary true outcome.

5.2 Individual level surrogacy: LRF reintroduction

Consider the continuous-continuous setting at the individual level. The LRF is based on the amount of information gained about the true outcome after accounting for the surrogate. Alonso et al. (2006) proposed using two models for continuous outcomes for each trial i:

$$T_{ij} = \mu_i + \beta_i Z_{ij} + \varepsilon_{T_{ij}} \quad 5.1$$

$$T_{ij} = \theta_{0_i} + \theta_{1_i} Z_{ij} + \theta_{2_i} S_{ij} + \varepsilon_{T|S_{ij}}, \quad 5.2$$

where: θ_{0_i} and μ_i are intercept parameters with and without adjustment for the surrogate; β_i is the treatment effect parameter for true outcome without adjustment for the surrogate; θ_{1_i} and θ_{2_i} are treatment and surrogate parameters for the model which adjusts for the surrogate. The difference in the amount of information on the true outcome gained from the surrogate is calculated via the difference in the log-likelihood between 5.1 and 5.2, which is formally expressed as G_i^2, for each trial i. L_0 is always the log-likelihood for the unsaturated model, in this case 5.1, and LL_1 for the saturated model, 5.2, for trial i. $G_i^2 = 2 * (LL_1 - LL_0)$. The LRF is then calculated:

$$LRF = R_h^2 = 1 - \frac{1}{N}\sum_i \exp\left(-\frac{G_i^2}{n_i}\right) \quad 5.3$$

R_h^2 can be interpreted as the amount of uncertainty in T that that can be explained by knowledge of S, for more details see section 3.3.1.

5.2.1 Individual level: ordinal-binary

This methodology is extended to the case of an ordinal surrogate and binary true outcome. I replaced the models of 5.1 and 5.2 with two generalised linear models, for each trial *i*, as the response variable is the binary true outcome.

$$logit[P(T_{ij} = 1)] = \mu_i + \beta_i Z_{ij} \quad 5.4$$

$$logit[P(T_{ij} = 1)] = \theta_{0_i} + \theta_{1_i} Z_{ij} + \theta_{2_i} S_{ij}, \quad 5.5$$

where the parameters in these models are directly comparable to those in the continuous setting above. In the case of discrete true outcomes the individual level was bounded above by a number less than one and had to be rescaled (Alonso and Molenberghs, 2007). This was done in exactly the same manner as the binary-ordinal setting, see section 3.3.2.

5.2.1.1 Modelling the ordinal surrogate explanatory variable

In this case the surrogate variable in 5.5 was ordinal. This could be modelled using one of two methods.

The first option was to model the ordinal variable as a quantitative variable, assuming that the underlying scale was continuous. This would be under the assumption that the observed ordinal variable was linearly related to an underlying continuum which was in turn linearly related, in this case through the logit link, to the response variable (Winship and Mare, 1984). This assumption is equivalent to the proportional odds assumption for the ordinal response variable of the binary-ordinal setting.

The alternative method was to model the ordinal variable as a factor using dummy variables. If there were seven categories in the ordinal outcome then six dummy variables were produced, each against a constrained reference category. This set up ignored the natural ordering of the ordinal variable and was un-parsimonious.

When the methods are compared there appears to be more draw backs to the dummy variable option. The fact that it ignored the ordering of the ordinal variable would lead to loss of information and its un-parsimonious nature may cause a lack of

degrees of freedom in the case of small trial sizes. Therefore, I modelled the ordinal explanatory variable as an interval scale variable and assumed that it was linearly related to the underlying continuum.

5.3 Trial level surrogacy: LRF reintroduction

In order to apply the LRF at the trial level a two stage approach is required. Again I first describe the case where continuous outcomes are used. At the first stage, two linear models for each trial *i* are required:

$$S_{ij} = \mu_{S_i} + \alpha_i Z_{ij} + \varepsilon_{ij} \quad 5.6$$

$$T_{ij} = \mu_{T_i} + \beta_i Z_{ij} + \varepsilon_{ij} \quad 5.7$$

An alternative way to achieve this is by using two models which incorporate all trials:

$$S_{ij} = \mu_{S_i} trial_{ij} + \alpha_i (trial_{ij} * Z_{ij}) + \varepsilon_{ij} \quad 5.8$$

$$T_{ij} = \mu_{T_i} trial_{ij} + \beta_i (trial_{ij} * Z_{ij}) + \varepsilon_{ij} \quad 5.9$$

In either case we return two variables, α_i and β_i, of the treatment effect estimates on the surrogate and true outcome respectively for each trial *i*. μ_{T_i} and μ_{S_i} are the intercepts for each trial *i* of the true outcome surrogate regressed on treatment respectively. At the second stage two further models are required: the intercept only model, 3.15; and treatment effect estimates of true outcome regressed on intercept and treatment effects of the surrogate for each trial *i*, 3.16.

$$\hat{\beta}_i = \gamma_3 + \varepsilon_i \quad 5.10$$

$$\hat{\beta}_i = \gamma_0 + \gamma_1 \hat{\mu}_{S_i} + \gamma_2 \hat{\alpha}_i + \varepsilon_i, \quad 5.11$$

where γ_0 and γ_3 are intercept parameters with and without adjustment for the surrogate; and γ_1 and γ_2 are the parameters for the surrogate intercept and treatment estimate variables. The difference in the 2*log-likelihood between these two models can then be calculated and the LRF applied as in 3.17.

$$LRF = \widehat{R_{ht}^2} = 1 - \exp\left(-\frac{G^2}{N_T}\right) \quad 5.12$$

The LRF is rescaled since it was found to be bounded above by a number less than one in the discrete case (Alonso and Molenberghs, 2007) this was done in exactly the same manner as the binary-ordinal setting, see section 3.3.2.

5.3.1 Trial level: ordinal-binary

In the ordinal-binary setting, the same procedure is followed as that for continuous outcomes. At the first stage of modelling, since we have an ordinal surrogate and a binary true outcome as response variables, a proportional odds model and a generalized linear model are required. These can again be used to calculate treatment effect estimates. Here I followed the approach used in the binary-ordinal setting and used full models incorporating all trials, in a similar manner to 5.8 and 5.9.

$$logit[P(S_{ij} \leq v)] = \mu_{S_v}^0 + \mu_{S_i} trial_j + \alpha_i (Z_{ij} * trial_j) \quad 5.13$$

$$logit[P(T_{ij})] = \mu_{T_i} trial_j + \beta_i (Z_{ij} * trial_j) \quad 5.14$$

Here, 1.14 is a proportional odds model, where, $v = 1, \ldots, V-1$, and V is the number of categories in the ordinal surrogate. In this case $\mu_{S_v}^0$ are the assumed fixed intercept cut points relating to the ordinal variable and μ_{S_i} are trial specific shifts of the set of intercepts (Burzykowski et al., 2004). $\mu_{S_V}^0$ and μ_{S_i} are the natural parameter estimates one would receive from the model under the assumption of proportional odds for all the explanatory variables. All other parameters are the same as those produced in the continuous-continuous setting.

At the second stage of modelling, models **3.15** and **3.16** can be fitted. Then the LRF can be calculated (using **3.17**) in a similar way as for the continuous-continuous case (using μ_{S_i}, α_i and β_i). Here, the trial specific shift of the set of intercepts, μ_{S_i}, can be used as a measure of the variability of the intercepts over trials, as advocated by Burzykowski et al. (2004) in the meta-analytical setting.

5.4 Confidence intervals: ordinal-binary

Confidence intervals for trial and individual level surrogacy in the ordinal-binary setting can be applied in exactly the same manner as the binary-ordinal setting, see section 3.5.

5.5 Separation: ordinal-binary

A solution to the issue of separation is to use the penalized likelihood technique of Firth (1993). To see more details of this method and the issues of separation see section 3.6.

The penalized likelihood technique can be applied in the ordinal-binary setting in much the same manner as in the binary-ordinal setting (section 3.6.4). The only difference is that in the ordinal-binary setting calculation of intercept variables is complicated. The publically available penalized likelihood technique software for ordinal outcomes cannot currently deal with large datasets that have hierarchical structures. Since this is the case, the penalized technique requires using separate models of surrogate and true outcome regressed on treatment for each trial i, as in 5.15 and 5.16.

$$logit[P(S_{ij} \leq v)] = \mu_{S_{v_i}} + \alpha_i Z_{ij} \qquad 5.15$$

$$logit[P(T_{ij} = 1)] = \mu_{T_i} + \beta_i Z_{ij}, \qquad 5.16$$

where, $v = 1, \ldots, V - 1$.

These separate models are equivalent to 5.6 and 5.7 in the continuous-continuous setting. In this case, the proportional odds model of the separate models (5.15) would return a separate intercept parameter for each cut-point of the ordinal outcome variable for each trial, $\mu_{S_{v_i}}$. All other parameters are equivalent to those in the models of the continuous setting (5.6 and 5.7). The use of such a set of variables is also suggested by Burzykowski et al. (2004) in the meta-analytical context. In the information theory context, the second stage model **3.16** can be replaced with the following model, using all the surrogate intercept variables of 5.15.

$$\hat{\beta}_i = \rho_0 + \rho_1 \hat{\mu}_{S_{1_i}} + \rho_2 \hat{\mu}_{S_{2_i}} + \cdots + \rho_{V-1} \hat{\mu}_{S_{V-1_i}} + \rho_S \hat{\alpha}_i + \varepsilon_i \quad 5.17$$

Here, $\hat{\mu}_{S_{1_i}}, \ldots, \hat{\mu}_{S_{V-1_i}}$ are the intercept estimates for each cut point modelled along with a treatment effects estimate variable, $\hat{\alpha}_i$. $\rho_0, \rho_1, \ldots, \rho_{V-1}, \rho_S$ are their corresponding parameter estimates respectively.

However, when there are only a few trials this model would lead to overfitting. In the case of only five trials, an ordinal variable with three categories and two cut-points leads to a model based on 5.17 with three explanatory variables; two intercept variables, $\hat{\mu}_{S_{1_i}}$ and $\hat{\mu}_{S_{2_i}}$, and treatment, $\hat{\alpha}_i$. This model would be based on five observations (the treatment estimates for the five trials) and result in issues of lack of degrees of freedom.

One way to resolve this would be to calculate the mean intercept value for each trial over all cut-points. This would provide one estimate of the intercept for each trial, $\widehat{\bar{\mu}_{S_{V_i}}}$. Then stage two models of trial level surrogacy would become:

$$\hat{\beta}_i = \gamma_3 + \varepsilon_i \quad 5.18$$

$$\hat{\beta}_i = \gamma_0 + \gamma_1 \widehat{\bar{\mu}_{S_{V_i}}} + \gamma_2 \hat{\alpha}_i + \varepsilon_i, \quad 5.19$$

where parameters are defined in the same manner as in the continuous case, see equations 3.15 and 3.16. The LRF could be applied to these models using **3.17** in the same manner as in the continuous-continuous case.

5.6 Conclusions: ordinal-binary

I have extended the information theory approach to the case of an ordinal surrogate and binary true outcome.

In addition to the theory discussed in the binary-ordinal setting, this has included: a discussion of the best way to model an ordinal surrogate explanatory variable for the models at the individual level; using a trial specific shift variable for calculating intercept estimates for use in trial level surrogacy; and a mean intercept variable to

do the same when separation occurs as the trial specific shift variable cannot be produced.

The trial specific shift method is more in keeping with the method used in other settings to represent intercept variations across trials than the mean intercept method. However, the mean intercept method was expected to return comparable results.

Chapter 6. Simulation study: for an ordinal surrogate and binary true outcome

In the previous chapter I discussed how the information theoretic measures R_h^2 and R_{ht}^2 for surrogate evaluation can be extended to the case of an ordinal surrogate and binary true outcome. In this chapter, I present a simulation study conducted to investigate the behaviour of these measures in the ordinal-binary setting under various scenarios.

I outline the setup of the simulation study, various practical considerations and present simulation results for both R_h^2 and R_{ht}^2.

As in previous chapters: S denotes a surrogate; T a true outcome; Z treatment; with i=1,..,N trials and j=1,..,n_i patients per trial.

6.1 Simulation study: set up

In the binary-ordinal setting I investigated approaches used in simulation studies for surrogate evaluation, see section 4.1.1.1. I concluded that the one presented by Burzykowski et al. (2005) would be best for my purposes. I adopted this set up for the simulation in the ordinal-binary setting.

First, a continuous S and T were simulated using a joint mixed model. Then these variables were categorised into an ordinal S and a binary T in the same manner as in the binary-ordinal setting only on the opposite outcomes. This will be further discussed in section 6.1.1.

The models used to simulate the study in the ordinal-binary setting were exactly the same as in the binary-ordinal setting. For details, see section 4.1.2.

The scenarios investigated in the simulation for the binary-ordinal setting are shown in Table 6.1. These included:

- varying the number and size of trials;
- varying the strength of surrogacy at the trial and individual levels;

- differing the strength of surrogacy at the trial and individual levels;
- examining both proportional and non-proportional odds in the ordinal surrogate outcome at the trial level;
- and examining linear and non-linear relationships in outcomes at the individual level.

The investigation of the assumptions (proportional odds and linear relationship) tested in the ordinal-binary setting were slightly different to those in the binary-ordinal setting. These scenarios will be discussed in section 6.1.2.

Factor varied under simulation study	Levels of factor
Number of trials	5, 10, 20 or 30
Number of patients per trial Small trial size Large trial size	 10, 20, 40 or 60 100, 150, 200 or 300
True surrogacy strength	R_h^2=1, 0.64 or 0.30 and R_{ht}^2=0.90 or 0.30
Agreement of surrogacy strength	
• Agree both strong or both weak	R_h^2= 0.64 & R_{ht}^2=0.90 Or R_h^2= 0.30 & R_{ht}^2=0.30
• Disagree one weak one strong	R_h^2= 0.64 & R_{ht}^2=0.30 Or R_h^2= 0.30 & R_{ht}^2=0.90
Trial level: Adherence to the proportional odds assumption	 Proportional odds Non proportional odds
Individual level: Linear relationship assumption	 Linear relationship Non-linear relationship

Table 6.1: *Scenarios investigated in the binary-ordinal simulation.*

In each scenario 250 simulations were conducted. The median R^2 values from the 250 simulations for each scenario are presented in the results section. This is to give some idea of the bias in the R^2 estimates. The variance of R^2 estimates in the 250 simulations are also presented to assess precision. In each scenario, the median lower and upper confidence bounds were determined, over the 250 individual confidence intervals for each simulated dataset. These give an indication of how well the confidence intervals perform under each scenario.

The practical set-up of the simulation was the same as the binary-ordinal setting in terms of the coding, saving and reported output, see section 4.1.3.3. A few aspects of the simulation differed in the ordinal-binary to the binary-ordinal setting or should be reintroduced. These are discussed below.

6.1.1 Loss of information

6.1.1.1 Loss of information: individual level

As in the binary-ordinal setting, the ordinal-binary setting was impacted by loss of information. This occurred because the simulated continuous S and T were categorised into discrete variables. This scenario was reflective of real life where discrete variables may represent underlying continua and therefore this setup was appropriate (see section 4.1.3.1 for more information).

As in the binary-ordinal setting we cannot determine the true strength of surrogacy in the observed ordinal-binary setting where this was set to a certain strength at the underlying continuous setting. See section 4.1.3.3.1 for more information.

To determine the ceiling in the ordinal-binary setting a third scenario was added to the simulation. Setting surrogacy to be 'perfect' (i.e. $R_h^2 = 1$) in the underlying continuous-continuous setting enabled me to determine empirically the highest value a surrogate could possibly take in the observed ordinal-binary setting.

6.1.1.2 Loss of information: trial level

Loss of information impacted trial level surrogacy in the binary-ordinal setting despite the fact that theoretically this should not occur, see section 4.2.2.1 for a discussion of this topic.

Theoretically, the two stage nature of the approach effectively split a large computationally difficult problem into separate components which were then solved individually. The solutions to each component were brought together to give an overall result for the whole problem. The splitting of the problem combined with loss of information due to categorisation has been shown to lead to loss of efficiency, (Molenberghs et al., 2011), (Taylor and Yu, 2002) and (Taylor et al., 2006). Theoretically, this loss of efficiency worsens as the number of components (trials) the problem is split into increases. This loss of efficiency should theoretically improve if the corresponding size of the components is large, the reasoning for this was thoroughly explained in section 4.2.2.1.2.

6.1.2 Assumptions

6.1.2.1 Proportional odds assumption: trial level

At the first stage of trial level surrogacy evaluation, a proportional odds model for the ordinal surrogate was applied, see section 5.3.1. Therefore, it was appropriate to assess whether deviations from the proportional odds assumption of the ordinal surrogate impacted results.

In the simulation study proportional odds were simulated in exactly the same manner as in the binary-ordinal setting, see section 4.1.3. As in the binary-ordinal setting, diversions from the proportional odds assumption were investigated by differing the categorisation of treatment groups. One treatment group had ordinal cut points based on the proportional odds scenario, see Table 6.2 column one. The other treatment group followed this rule for five of the categories but deviated for two, where one group was based on a much larger range of quantiles than the other, see Table 6.2 column two. This was based on the setup of the simulation for the binary-ordinal setting, see section 4.1.3.2.

To investigate the median simulated true surrogate proportional and non-proportional odds ratios at each cut point of the ordinal surrogate see Table 6.3, this simulation was conducted in the same was as for the true outcome in the binary-ordinal setting see section 4.1.3.2. These simulated median odds ratios demonstrate that in the proportional odds setting the odds are indeed proportional at each cut point of the ordinal surrogate, and in the non-proportional odds setting the odds are proportional at all but one of the cut point as anticipated.

	Column 1 Treatment group one		Column 2 Treatment group two	
Ordinal category	Quantiles continuous	Difference in quantiles	Quantiles continuous	Difference in quantiles
1	≤0.143	0.143	≤0.25	0.25
2	0.143-0.286	0.143	0.25-0.286	0.036
3	0.286-0.429	0.143	0.286-0.429	0.143
4	0.429-0.571	0.143	0.429-0.571	0.143
5	0.571-0.714	0.143	0.571-0.714	0.143
6	0.714-0.857	0.143	0.714-0.857	0.143
7	≥0.857	0.143	≥0.857	0.143

Table 6.2: *Set up of non-proportional odds scenario*.

Cut point ordinal surrogate outcome	Proportional odds scenario	Non proportional odds scenario
1	0.9749	0.4926
2	0.9637	0.9892
3	0.9635	0.9776
4	0.9652	0.9767
5	0.9700	0.9785
6	0.9574	0.9749

Table 6.3: *Simulated odds were based on 1000 runs of the study set up with 30 trials, 300 patients, and strong surrogacy at both levels, the median odds ratios over all simulated cases are given for each cut point.*

6.1.2.2 Proportional odds assumption: individual level

In the ordinal-binary setting at the individual level, the models for assessing surrogacy were based on generalised linear models for the binary response variable true outcome, see section 5.2.1. Unlike in the binary-ordinal setting, no proportional odds models were used, therefore it was not necessary to investigate the non-proportional odds assumption for individual level surrogacy.

6.1.2.3 Linear relationship assumption: individual level

As discussed in section 5.2.1.1, one of the models of individual level surrogacy incorporated the ordinal surrogate as a quantitative explanatory variable. This is under the assumption that the observed ordinal variable is linearly related to an underlying continuum which is linearly related, in this case through the logit link, to the response variable (Winship and Mare, 1984). This is the equivalent assumption to the proportional odds assumption for an ordinal explanatory variable. Therefore, I used the non-proportional coding to test this assumption: a justification of this statement follows.

In the simulation, a simulated continuous S was linearly related to the continuous T (the response variable). However, if S was categorised using the setup for the non-proportional scenario, two of the categories of the ordinal S would not be linearly related to the simulated continuous S. Therefore, this assumption would be invalid. Hence, I used the coding of the non-proportional odds scenario to investigate the linear relationship assumption imposed on models at the individual level.

6.1.3 Trial level: separation

Separation is a common occurrence for discrete outcomes and under the information theory approach leads to severe bias at the trial level, see section 3.6. Two techniques were compared to assess the best for dealing with separation: 1) the penalized likelihood technique; and 2) the technique which removed trials where separation occurred from analysis.

The calculation of the trial specific intercept values for use in stage two models is more complicated in the ordinal-binary compared to the binary-ordinal setting. This

is because of the measures returned from models based on ordinal response variables (in this case the surrogate), see section 5.3.1. In the case of the penalized likelihood technique a 'mean intercept' method was adopted. The trial removal technique uses a 'trial specific shift' method, this method is more in keeping with the approach used in the binary-ordinal setting. Comparison of the penalized likelihood and trial removal techniques served as an indicator of how well the intercept methods worked.

6.1.4 Set up: conclusions

The setup of the simulation in the ordinal-binary setting has been outlined. This was similar to that of the binary-ordinal setting. Except that the proportional odds assumption was not applied at the individual level and the equivalent linear relationship assumption was required. Additionally, calculation of intercept variables was more complicated in the ordinal-binary setting; results were examined to identify if any issues arose as a result of this.

6.2 Simulation study: results

In this section results of the simulation study for the individual level and then the trial level surrogacy is presented. Various issues and biases in the results are outlined in Table 6.4.

Issues of loss of information (at the individual level), underestimation and separation (at the trial level) highlighted in rows one, two and four of Table 6.4, respectively have been discussed earlier in this Chapter, see sections 4.1.3.3.1, 4.2.2.1.2 and 4.2.2.2.1. The problem of overestimation, also shown in row three of this table, was discussed for the binary-ordinal setting in section 4.2.2.2. Issues described in Table 6.4 are present in the results of the ordinal-binary setting and will be discussed further in the sections indicated in Table 6.4. Some of these issues conflict with each other making description of results complicated. Therefore, this table should be considered to aid understanding.

	Issue	**Scenarios affected**	**Reason**	**Discussed in section(s):**
$\boldsymbol{R_h^2}$	R_h^2 estimates substantially lower than true value	All scenarios	Loss of information due to categorisation	6.1.1.1 and 6.2.1.5
$\boldsymbol{R_{ht}^2}$	Underestimation	Worse for: • Larger numbers of trials • Strong surrogacy • Small trial sizes	Inefficiency in estimation due to categorisation	6.1.1.2 and 6.2.2.1
	Overestimation	Worse for: • Small numbers of trials • Weak surrogacy	Model fitting issue	6.2.2.2
	Separation	All scenarios	Zero cells in trial crosstabs Resolved through use of Firth (1993) technique	6.1.3 and 4.2.2.5

Table 6.4: *Issues present in simulation study results.*

Preliminary findings prompted the secondary investigations of interest on:

- the ceiling in R_h^2;
- and the benefit of the penalized likelihood technique for R_{ht}^2 when separation occurs.

The following summary results are given: the median R^2 for a given scenario; the IQRs of R^2 the 100% coverage of the confidence intervals and their median upper and lower bounds for each scenario. The tables in the remainder of this chapter show results for a subsection of the trial sizes investigated. Unless otherwise stated, results for scenarios not included in tables were consistent with those shown. Full tables for every setting are provided in Appendix B.

6.2.1 Results: individual level surrogacy

Results for the individual level surrogacy are described in this section. I describe the following scenarios:

- where surrogacy was set to be strong at both trial and individual level;
- where this was weak;
- where the strength of surrogacy differed at trial and individual levels;
- where there were deviations from the linear relationship assumption.

I also present the results of the ceiling investigation and compare these to the binary-ordinal setting.

6.2.1.1 Individual level surrogacy R^2_h: strong surrogacy

Surrogacy was set to be strong at both levels, R_h^2=0.64 and R_{ht}^2=0.90, in the underlying continuous setting. As previously mentioned, see section 6.1.1.1, loss of information means that the observed results for the ordinal-binary setting were expected to be lower than in the underlying continuum.

The results in Table 6.5 show that as the number and sizes of trials increased the value of surrogacy in the observed ordinal-binary setting converged to around 0.39. The impact of trial size was greater than that off number of trials, however all values returned were close to 0.39. The IQRs of R_h^2 were much narrower as the number and size of trials increased. The coverage of the confidence intervals was 100% in nearly all scenarios. The median confidence intervals covered nearly the whole parameter

space where there were small trial sizes but these decreased as the size of trials increased.

Number of trials	Trial size	R_h^2	IQR R_h^2	% Cover CIs	Median CI lower	Median CI upper
5	10	0.452	0.177	100%	0.051	0.999
5	60	0.410	0.106	100%	0.168	0.687
5	100	0.414	0.096	100%	0.214	0.631
5	300	0.405	0.068	98%	0.287	0.531
10	10	0.446	0.140	100%	0.054	0.989
10	60	0.399	0.076	100%	0.164	0.674
10	100	0.395	0.058	100%	0.204	0.610
10	300	0.397	0.054	100%	0.279	0.522
20	10	0.420	0.104	100%	0.049	0.964
20	60	0.396	0.058	100%	0.165	0.667
20	100	0.393	0.045	100%	0.205	0.607
20	300	0.393	0.037	100%	0.277	0.517
30	10	0.424	0.074	100%	0.051	0.964
30	60	0.397	0.040	100%	0.163	0.669
30	100	0.393	0.033	100%	0.203	0.605
30	300	0.389	0.033	100%	0.273	0.513

Table 6.5: *Simulation study: Median R_h^2 estimates based on 250 simulations for each scenario, where true values set to: R_h^2 =0.64 and R_{ht}^2 =0.90.*

6.2.1.2 Individual level surrogacy R^2_h: weak surrogacy

The findings for strong surrogacy were mirrored in the case of weak surrogacy, where surrogacy was set to R_h^2 =0.30 and R_{ht}^2=0.30 in the underlying continuous setting.

The results in Table 6.6 show that as the number and size of trials increased the value of surrogacy in the observed ordinal-binary setting converged to about 0.17. The impact of trial sizes was much greater than that off the number of trials. In the case of small trial sizes, estimates were as large as 0.282.

Number of trials	Trial size	R_h^2	IQR R_h^2	% Cover CIs	Median CI lower	Median CI upper
5	10	0.296	0.184	100%	0.017	0.868
5	60	0.195	0.069	100%	0.040	0.445
5	100	0.182	0.055	100%	0.054	0.375
5	300	0.181	0.038	100%	0.095	0.287
10	10	0.266	0.115	100%	0.019	0.851
10	60	0.187	0.048	100%	0.038	0.434
10	100	0.180	0.039	100%	0.054	0.368
10	300	0.173	0.030	100%	0.090	0.278
20	10	0.279	0.081	100%	0.023	0.855
20	60	0.182	0.030	100%	0.038	0.429
20	100	0.177	0.033	100%	0.053	0.362
20	300	0.175	0.019	100%	0.092	0.280
30	10	0.280	0.067	100%	0.022	0.854
30	60	0.184	0.026	100%	0.039	0.428
30	100	0.179	0.024	100%	0.054	0.366
30	300	0.172	0.017	100%	0.089	0.275

Table 6.6: *Simulation study: Median R_h^2 estimates based on 250 simulations for each scenario, where true values set to: R_h^2 =0.30 and R_{ht}^2 =0.30.*

6.2.1.3 Individual level surrogacy R^2_h: differing strengths of surrogacy

The results of R_h^2 for discordant strengths of surrogacy were much the same as those for agreement in surrogacy strength, see Table 6.7. Those for discordant strengths were marginally lower than for agreement. The coverages and the IQRs were much the same.

Number of trials	Trial size	Surrogacy strong both levels			Surrogacy strong R_h^2, weak R_{ht}^2		
		R_h^2 =0.90	IQR R_h^2	% Cover	R_h^2 =0.90	IQR R_h^2	% Cover
5	10	0.452	0.177	100%	0.443	0.206	100%
5	60	0.410	0.106	100%	0.415	0.088	100%
5	100	0.414	0.096	100%	0.409	0.089	100%
5	300	0.405	0.068	98%	0.403	0.073	96%
10	10	0.446	0.140	100%	0.433	0.138	100%
10	60	0.399	0.076	100%	0.405	0.072	100%
10	100	0.395	0.058	100%	0.397	0.056	100%
10	300	0.397	0.054	100%	0.388	0.055	100%
20	10	0.420	0.104	100%	0.429	0.087	100%
20	60	0.396	0.058	100%	0.400	0.044	100%
20	100	0.393	0.045	100%	0.389	0.047	100%
20	300	0.393	0.037	100%	0.389	0.038	100%
30	10	0.424	0.074	100%	0.425	0.079	100%
30	60	0.397	0.040	100%	0.395	0.042	100%
30	100	0.393	0.033	100%	0.393	0.034	100%
30	300	0.389	0.033	100%	0.384	0.031	100%

Table 6.7: *Simulation study: Median R_h^2 estimates based on 250 simulations for each scenario, where true values set to: R_h^2 =0.64 and R_{ht}^2 =0.90 or 0.30.*

6.2.1.4 Individual level surrogacy R^2_h: linear relationship assumption

There was very little impact of the divergence from the linear relationship assumption on estimates of surrogacy at the individual level, see Table 6.8. The results where the linearity assumption was not valid were uniformly slightly lower than those where the assumption held, regardless of the number or size of trials. However, this was only a minor difference in results.

Number of trials	Trial size	Linear relationship assumption valid			Linear relationship assumption invalid		
		R_h^2 =0.90	IQR R_h^2	% Cover	R_h^2 =0.90	IQR R_h^2	% Cover
5	10	0.452	0.177	100%	0.449	0.204	100%
5	60	0.410	0.106	100%	0.416	0.097	100%
5	100	0.414	0.096	100%	0.406	0.081	100%
5	300	0.405	0.068	98%	0.399	0.070	97%
10	10	0.446	0.140	100%	0.437	0.158	100%
10	60	0.399	0.076	100%	0.398	0.067	100%
10	100	0.395	0.058	100%	0.403	0.060	100%
10	300	0.397	0.054	100%	0.391	0.048	100%
20	10	0.420	0.104	100%	0.414	0.102	100%
20	60	0.396	0.058	100%	0.392	0.052	100%
20	100	0.393	0.045	100%	0.390	0.044	100%
20	300	0.393	0.037	100%	0.383	0.043	100%
30	10	0.424	0.074	100%	0.423	0.076	100%
30	60	0.397	0.040	100%	0.393	0.038	100%
30	100	0.393	0.033	100%	0.386	0.036	100%
30	300	0.389	0.033	100%	0.382	0.028	100%

Table 6.8: *Simulation study: Median R_h^2 estimates based on 250 simulations for each scenario, where true values set to: R_h^2 =0.64 and R_{ht}^2 =0.90.. Comparing results for non-linear relationship scenario.*

6.2.1.5 Individual level surrogacy R^2_h: ceiling effect

The results showed that the strength of individual level surrogacy in the ordinal-binary setting compared to the underlying continuum was much lower due to loss of information. I investigated R_h^2 results in the observed ordinal-binary setting when the strength of surrogacy was 'perfect' at the individual level in the underlying continuous setting. This was done to see if there was a ceiling on surrogacy strength in the ordinal-binary setting.

As can be seen in Table 6.9 the value of surrogacy of the observed ordinal surrogate converges to around 0.70 as numbers and sizes of trials increases. In the case of

perfect surrogacy at the underlying continuum, this suggests that the strongest an ordinal surrogate for a binary true outcome could hope to be is around 0.70.

Number of trials	Trial size	R_h^2	IQR R_h^2	% Cover CIs	lower 95% CI	upper 95% CI
5	10	0.620	0.220	100%	0.107	1
5	60	0.733	0.140	97%	0.461	0.969
5	100	0.726	0.136	96%	0.516	0.916
5	300	0.733	0.135	75%	0.611	0.847
10	10	0.615	0.133	100%	0.113	1
10	60	0.723	0.098	100%	0.460	0.955
10	100	0.718	0.088	99%	0.510	0.904
10	300	0.717	0.096	90%	0.596	0.828
20	10	0.603	0.093	100%	0.108	1
20	60	0.704	0.066	100%	0.441	0.937
20	100	0.707	0.071	100%	0.502	0.891
20	300	0.707	0.071	100%	0.586	0.819
30	10	0.595	0.078	100%	0.107	1
30	60	0.697	0.053	100%	0.436	0.932
30	100	0.701	0.049	100%	0.496	0.887
30	300	0.704	0.054	100%	0.587	0.816

Table 6.9 *Simulation study: Median R_h^2 estimates based on 250 simulations for each scenario, where true values set to: R_h^2=1 and R_{ht}^2=0.90.*

As in the binary-ordinal setting the confidence intervals for the case of 'perfect' surrogacy seem to be behaving differently to previous scenarios, see section 4.2.1.5. As in the binary-ordinal setting the scenario of 'perfect' surrogacy presented here is a very particular case of surrogacy that is unlikely to be seen in practice.

6.2.1.6 Individual level surrogacy R^2_h: comparison to binary-ordinal setting

Since one outcome was dichotomised and one categorised in both the binary-ordinal and ordinal-binary settings one might have expected to see similar ceiling results. However, the ceiling in the ordinal-binary setting was much higher, around 0.70,

than in the binary-ordinal setting, around 0.47. Observed results for all strengths of surrogacy at the individual level were lower for the binary-ordinal setting.

Surrogacy reflects the amount of information on the true outcome that can be provided by the surrogate. Therefore, it makes sense that, the amount of information retained in the surrogate was most pertinent to the level of the ceiling. The ceiling was lowest when the surrogate was binary (i.e. where more information was removed from the underlying continuum compared to an ordinal outcome). This is regardless of the fact that the ordinal-binary setting had a binary true outcome.

6.2.1.7 Individual level surrogacy R^2_h: conclusions

R^2_h estimates converged to around 0.39 when this was set to 0.64 at the underlying continuum, and 0.17 where this was set to 0.30 at the underlying continuum. Smaller trials gave larger R^2_h estimates and wider IQRs. The number of trials had limited impact. Differing strengths of surrogacy at trial and individual levels and non-adherence to the linear relationship assumption had little impact on R^2_h estimation.

The coverages of the confidence intervals were 100% in nearly all scenarios, that are likely to be seen in practice, suggesting these are conservative. Although, the median confidence interval results suggest the intervals were sensible and by no means covered the whole parameter space.

The median R^2_h value converged to about 0.70 where surrogacy was set to one in the underlying continuous-continuous setting. This corresponds to a large loss of information. This ceiling effect was much higher than that observed in the binary-ordinal setting.

6.2.2 Results: trial level surrogacy

In this setting I describe the results at the trial level:

- where surrogacy was strong at both the trial and individual levels;
- where surrogacy was weak at both levels;
- where the strength of surrogacy disagreed at trial and individual levels;
- where non-proportional odds of the ordinal surrogate were present.

- Finally, I compare the penalised likelihood technique to deal with separation to the technique where trials with separation were removed from analysis.

The underestimation and overestimation issues presented in Table 6.4 affected the results at the trial level.

6.2.2.1 Trial level surrogacy R^2_{ht}: strong surrogacy

In Table 6.10, the results for trial level surrogacy when surrogacy was strong at both trial and individual levels and R^2_{ht} was simulated to be 0.90 are shown.

Number of trials	Trial size	R^2_{ht}	IQR R^2_{ht}	% Cover CIs	Median CI lower	Median CI upper
5	10	0.760	0.376	92%	0.097	0.986
5	60	0.907	0.165	95%	0.358	0.997
5	100	0.923	0.172	94%	0.408	0.998
5	300	0.944	0.116	93%	0.489	0.999
10	10	0.588	0.324	55%	0.097	0.913
10	60	0.823	0.188	92%	0.385	0.976
10	100	0.860	0.143	93%	0.457	0.983
10	300	0.888	0.104	98%	0.523	0.988
20	10	0.487	0.204	14%	0.133	0.793
20	60	0.771	0.142	75%	0.452	0.935
20	100	0.820	0.111	89%	0.531	0.953
20	300	0.862	0.077	97%	0.610	0.967
30	10	0.480	0.197	3%	0.184	0.744
30	60	0.760	0.103	61%	0.503	0.910
30	100	0.798	0.081	80%	0.560	0.928
30	300	0.854	0.071	96%	0.653	0.952

Table 6.10: *Simulation study: Median R^2_{ht} estimates based on 250 simulations for each scenario, in the binary-ordinal setting where true values set to: R^2_h =0.64 and R^2_{ht} =0.90.*

Estimation improved as the size of trials increased. Results showed overestimation for data sets containing five trials. However, as the number of trials increased they

began to show underestimation. This was worst for the largest trial scenario of 30 trials; for example, the median R_{ht}^2=0.859for 300 patients (0.90 was expected). The IQRs of R_{ht}^2 and coverage of confidence intervals improved as size and numbers of trials increased. However, the coverage was very poor for smaller numbers of patients per trial which worsened as the number of trials increased, this was probably due to the relative bias in R_{ht}^2results for these scenarios.

6.2.2.1.1 Strong surrogacy: comparison to continuous-continuous setting

In Table 6.11, the results of strong surrogacy for the ordinal-binary scenario compared to the underlying continuous-continuous setting are presented. The underestimation witnessed in the ordinal-binary setting was not present at the underlying continuum when larger numbers of patients per trial were available

		Ordinal-binary setting		Continuous-continuous setting	
Number of trials	Trial size	R^2_{ht}	IQR R^2_{ht}	R^2_{ht}	IQR R^2_{ht}
5	10	0.760	0.376	0.931	0.165
5	60	0.907	0.165	0.950	0.103
5	100	0.923	0.172	0.953	0.095
5	300	0.944	0.116	0.959	0.069
10	10	0.588	0.324	0.840	0.137
10	60	0.823	0.188	0.909	0.096
10	100	0.860	0.143	0.916	0.072
10	300	0.888	0.104	0.919	0.068
20	10	0.487	0.204	0.835	0.100
20	60	0.771	0.142	0.900	0.065
20	100	0.820	0.111	0.902	0.054
20	300	0.862	0.077	0.910	0.058
30	10	0.480	0.197	0.826	0.078
30	60	0.760	0.103	0.895	0.053
30	100	0.798	0.081	0.900	0.050
30	300	0.854	0.071	0.902	0.054

Table 6.11: *Simulation study: Median R^2_{ht} estimates based on 250 simulations for each scenario, in binary-ordinal and continuous-continuous setting where true values set to: R^2_h =0.64 and R^2_{ht} =0.90.*

6.2.2.1.2 Underestimation R^2_{ht}: strong surrogacy

I outlined that categorising S and T into binary and ordinal outcomes led to loss of efficiency in calculating results at the trial level in section 6.1.1.2.

Proof of this theory of inefficiency was obtained by considering an additional scenario in the simulation study. Consider Table 6.12, 3000 patient scenarios for each trial size are presented. These results are much closer to the true strength of surrogacy and indicate that it was indeed inefficiency that caused the underestimation in results.

		binary-ordinal	
Number of trials	Trial size	R^2_{ht}	IQR R^2_{ht}
5	3000	0.956	0.071
10	3000	0.920	0.050
20	3000	0.891	0.035
30	3000	0.883	0.024

Table 6.12 *Additional simulation scenarios: Median R^2_{ht} estimates based on 250 simulations for each scenario, where R^2_h=0.64 and R^2_{ht}=0.90 and trial size was 3000.*

6.2.2.1.3 Overestimation R^2_{ht}: strong surrogacy

There was some evidence of overestimation in Table 6.10 for small number of trials. This issue was referenced in row three of Table 6.4 and will be discussed in section 6.2.2.2.2.

6.2.2.2 Trial level surrogacy R^2_{ht}: weak surrogacy

Here I discuss the case where surrogacy was weak at both levels, $R^2_h = R^2_{ht}$=0.30.

The results showed that as the number of trials decreased, overestimation worsened, see Table 6.13. Where there were five trials and 300 patients', trial level surrogacy was 0.663, indicating a moderately good surrogate in contrast to the true strength of surrogacy of 0.30. The IQRs were wider than the strong surrogacy scenario, for comparison see Table 6.11, but still showed improvement as the number and size of trials increased. Confidence intervals showed better coverage as the number of trials increased but did not differ greatly by increased trial size.

Number of trials	Trial size	R^2_{ht}	IQR R^2_{ht}	% cover CIs	Median CI lower	Median CI upper
5	10	0.619	0.431	84%	0.021	0.968
5	60	0.596	0.503	80%	0.016	0.965
5	100	0.672	0.506	83%	0.038	0.976
5	300	0.666	0.391	80%	0.036	0.975
10	10	0.285	0.350	96%	0.000	0.763
10	60	0.385	0.353	98%	0.011	0.823
10	100	0.399	0.357	95%	0.013	0.831
10	300	0.424	0.326	97%	0.019	0.844
20	10	0.225	0.232	94%	0.007	0.589
20	60	0.314	0.198	96%	0.030	0.669
20	100	0.324	0.271	98%	0.035	0.678
20	300	0.284	0.249	96%	0.020	0.644
30	10	0.196	0.155	90%	0.012	0.495
30	60	0.285	0.173	95%	0.048	0.584
30	100	0.297	0.199	98%	0.054	0.596
30	300	0.321	0.190	96%	0.068	0.618

Table 6.13: *Simulation study: median R^2_{ht} estimates based on 250 simulations for each scenario, in the binary-ordinal setting where true values set to: R^2_h =0.30 and R^2_{ht} =0.30.*

6.2.2.2.1 Comparison to underlying continuous-continuous setting: weak surrogacy

Results for the ordinal-binary setting were compared to the underlying continuous-continuous setting, shown in Table 6.14. The pattern of overestimation was consistent at both the observed ordinal-binary and underlying continuous-continuous settings. Results in the continuous-continuous setting showed overestimation in every setting. However, results for the ordinal-binary setting were consistently lower than the underlying continuous-continuous setting, the difference being greater for smaller trial sizes.

		Ordinal-binary setting		Continuous-continuous setting	
Number of trials	Trial size	R^2_{ht}	IQR R^2_{ht}	R^2_{ht}	IQR R^2_{ht}
5	10	0.619	0.431	0.600	0.453
5	60	0.596	0.503	0.659	0.399
5	100	0.672	0.506	0.676	0.539
5	300	0.666	0.391	0.683	0.440
10	10	0.285	0.350	0.438	0.298
10	60	0.385	0.353	0.446	0.347
10	100	0.399	0.357	0.402	0.376
10	300	0.424	0.326	0.429	0.324
20	10	0.225	0.232	0.349	0.212
20	60	0.314	0.198	0.353	0.199
20	100	0.324	0.271	0.327	0.213
20	300	0.284	0.249	0.355	0.222
30	10	0.196	0.155	0.326	0.172
30	60	0.285	0.173	0.329	0.179
30	100	0.297	0.199	0.358	0.210
30	300	0.321	0.190	0.336	0.204

Table 6.14: *Median R^2_{ht} estimates based on 250 simulations for each scenario, in continuous-continuous setting where true values set to: R^2_h =0.30 and R^2_{ht} =0.30.*

6.2.2.2.2 Overestimation R^2_{ht}: weak surrogacy

The overestimation witnessed in both the observed ordinal-binary and underlying continuous-continuous settings occurred because of overfitting in second stage models where the number of trials was small. The reason for this was thoroughly discussed in the binary-ordinal setting in section 4.2.2.2.1.

6.2.2.2.3 Underestimation R^2_{ht}: weak surrogacy

There was also evidence of underestimation when surrogacy was weak. Notice that the results in the ordinal-binary setting were consistently lower than in the underlying continuous-continuous setting. This was more noticeable as the size of trials decreased. The reason for this was again because of inefficiency due to loss of

information and a two stage approach as described in section 6.2.2.1 also see Table 6.4.

6.2.2.3 Trial level surrogacy R^2_{ht}: differing strengths of surrogacy

R^2_{ht} estimation was little affected where there was strong surrogacy at the trial level but weak surrogacy at the individual level compared to where this was strong at both levels, see Table 6.15. The estimates were lower for the discordant strengths scenario when there were few patients per trial. Confidence interval coverages were also poorer. Similar discrepant results were found when surrogacy was strong at the individual level and weak at trial levels, see Appendix B, Table B4.

Number of trials	Trial size	Surrogacy strong both levels			Surrogacy strong $\boldsymbol{R^2_{ht}}$, weak $\boldsymbol{R^2_h}$		
		R^2_{ht} =0.90	IQR R^2_{ht}	% cover	R^2_{ht} =0.90	IQR R^2_{ht}	% cover
5	10	0.760	0.376	92%	0.703	0.335	89%
5	60	0.907	0.165	95%	0.895	0.187	94%
5	100	0.923	0.172	94%	0.920	0.171	98%
5	300	0.944	0.116	93%	0.947	0.112	95%
10	10	0.588	0.324	55%	0.504	0.367	44%
10	60	0.823	0.188	92%	0.789	0.195	86%
10	100	0.860	0.143	93%	0.837	0.154	94%
10	300	0.888	0.104	98%	0.881	0.101	98%
20	10	0.487	0.204	14%	0.436	0.235	4%
20	60	0.771	0.142	75%	0.746	0.129	71%
20	100	0.820	0.111	89%	0.790	0.126	85%
20	300	0.862	0.077	97%	0.864	0.084	95%
30	10	0.480	0.197	3%	0.388	0.221	1%
30	60	0.760	0.103	61%	0.725	0.118	46%
30	100	0.798	0.081	80%	0.790	0.102	72%
30	300	0.854	0.071	96%	0.851	0.079	92%

Table 6.15: *Simulation study: Median R^2_{ht} estimates based on 250 simulations for each scenario, where true values set to: R^2_{ht} =0.90 and R^2_h =0.64 or 0.30.*

6.2.2.4 Trial level surrogacy R^2_{ht}: non-proportional odds

Slight differences in the estimates of surrogacy at the trial level were observed due to the divergence from the proportional odds assumption, see Table 6.16. R^2_{ht} estimates were marginally lower when the odds were not proportional and trial sizes were large and marginally larger where trial sizes were small, however these differences were minor. The results did not vary according to the number of trials and the coverage of the confidence intervals were generally comparable.

Number of trials	Trial size	Proportional			Non proportional		
		R^2_{ht} =0.90	IQR R^2_{ht}	% cover CIs	R^2_{ht} =0.90	IQR R^2_{ht}	% cover CIs
5	10	0.760	0.376	92%	0.758	0.401	85%
5	60	0.907	0.165	95%	0.905	0.185	94%
5	100	0.923	0.172	94%	0.925	0.149	96%
5	300	0.944	0.116	93%	0.931	0.111	96%
10	10	0.588	0.324	55%	0.562	0.299	52%
10	60	0.823	0.188	92%	0.808	0.202	90%
10	100	0.860	0.143	93%	0.828	0.150	94%
10	300	0.888	0.104	98%	0.885	0.099	99%
20	10	0.487	0.204	14%	0.505	0.196	11%
20	60	0.771	0.142	75%	0.756	0.156	70%
20	100	0.820	0.111	89%	0.805	0.131	83%
20	300	0.862	0.077	97%	0.859	0.088	96%
30	10	0.480	0.197	3%	0.495	0.195	2%
30	60	0.760	0.103	61%	0.756	0.112	56%
30	100	0.798	0.081	80%	0.797	0.097	76%
30	300	0.854	0.071	96%	0.852	0.074	91%

Table 6.16: S*imulation study: Median R^2_{ht} estimates based on 250 simulations for each scenario, where true values set to: R^2_{ht} =0.90 and R^2_h =0.64. Comparing results for proportional odds and non-proportional odds.*

6.2.2.5 Trial level surrogacy R^2_{ht}: dealing with separation

6.2.2.5.1 Occurrence of separation: trial level

The percentage of separations across scenarios for binary outcomes in the simulation ranged from 3.2% to 79.9%. The average was 23.4%. Separation in the ordinal case was recorded separately for complete and quasi-complete separation. Ordinal complete separation ranged from 0% to 14.4% with an average of 3.4%. Quasi-complete separation ranged from 0% to 21.9% with an average of 1.8%. Hence, separation occurred frequently under a random simulation process in the ordinal-binary setting and led to bias in R^2_{ht} which needed to be resolved.

6.2.2.5.2 Comparison of penalized likelihood and trial removal techniques: separation trial level

Two techniques were applied to deal with separation: the penalized likelihood technique and the trial removal technique (here trials where separation occurred were removed from analysis), see section 4.1.3.3.4.

In Table 6.17 the penalized likelihood technique was compared to the removal of trials technique for strong surrogacy at both levels. The results were roughly comparable across settings with the exception of small trial sizes where separation occurs most frequently. Comparison of these two sets of results was not straightforward because of the impact of removing different numbers of trials within a particular scenario in the trial removal technique.

Comparing the techniques for ten patients per trial, estimation was better for larger numbers of trials for the penalised technique and better for the trial removal technique for small numbers of trials. However, this does not take into account the fact that up to 92% of the simulations were rejected under the removal technique because less than three trials were available for analysis. Furthermore, when simulations were not rejected often a lot fewer than the stated number of trials were available for analysis. I found that as the number of trials removed increased so too did the value of R^2_{ht} in the removal technique. This was due to overestimation because of overfitting which occurs when there are only a few trials, outlined in Table 6.4 and section 6.1.1.2. Therefore, R^2_{ht} estimates for the trial removal

technique were artificially high where there were small numbers of trials and were not a result of more effective estimation compared to the penalized likelihood technique.

Number of trials	Trial size	Penalized likelihood technique		Trial removal technique			
		R^2_{ht} =0.90	IQR R^2_{ht}	R^2_{ht} =0.90	IQR R^2_{ht}	% failures	Median trial No.
5	10	0.760	0.376	0.823	0.423	92.0%	4
5	60	0.907	0.165	0.916	0.167	13.5%	5
5	100	0.923	0.172	0.925	0.150	4.9%	5
5	300	0.944	0.116	0.946	0.127	0.4%	5
10	10	0.588	0.324	0.674	0.496	48.7%	5
10	60	0.823	0.188	0.829	0.245	0.0%	9
10	100	0.860	0.143	0.859	0.169	0.0%	9
10	300	0.888	0.104	0.898	0.109	0.0%	10
20	10	0.487	0.204	0.431	0.519	7.7%	7
20	60	0.771	0.142	0.780	0.215	0.0%	17
20	100	0.820	0.111	0.835	0.113	0.0%	18
20	300	0.862	0.077	0.878	0.084	0.0%	19
30	10	0.480	0.197	0.284	0.414	0.8%	10
30	60	0.760	0.103	0.791	0.129	0.0%	26
30	100	0.798	0.081	0.826	0.101	0.0%	27
30	300	0.854	0.071	0.865	0.061	0.0%	29

Table 6.17: *Simulation study: Median R^2_{ht} estimates based on 250 simulations for each scenario, where true values set to: R^2_{ht} =0.90 and R^2_h =0.64. Comparing penalized likelihood technique against trial removal technique (trial removal technique results include the % of time the calculation of R^2_{ht} was not possible and the median number of trials available for analysis when it was).*

This finding was supported by the fact that the IQRs for the penalised likelihood technique were much narrower than the trial removal technique. The penalised likelihood technique gave consistent estimation regardless of the number of

occurrences of separation. The penalized likelihood technique was far superior to a trial removal technique in terms of estimation, occurrence of modelling errors and information retention. It is a useful addition to the information theory approach in the ordinal-binary setting.

6.2.2.5.3 Comparison of penalized likelihood and trial removal techniques: intercept calculation

In order to calculate surrogate intercept variables for second stage models the: penalized likelihood technique used a mean intercept method; and the trial removal technique used a trial specific shift method. The results for the penalized likelihood and trial removal techniques were much the same for larger trial sizes where the occurrences of separation were not as frequent. This suggested that the means of calculating the surrogate intercept variable in either case had not imposed bias in results and were fit for purpose.

6.2.2.6 Trial level surrogacy R^2_{ht}: comparison to binary-ordinal setting

Comparisons of trial level surrogacy between the ordinal-binary and binary-ordinal settings showed very few differences.

There was slightly greater underestimation via the penalised likelihood technique for small trial sizes in the ordinal-binary setting, see Table 6.17 and Table 4.17. However, this difference in underestimation in techniques was only minor.

6.2.3 Trial level surrogacy R^2_{ht}: conclusions

The results in the ordinal-binary setting showed that trial level surrogacy worked well in general. Coverage of the confidence intervals were good when there was little bias in R^2_{ht} results.

However, three issues affected results: underestimation; overestimation; and separation. These issues were all related to the models of the two stages of the information theory approach. Underestimation and bias imposed by separation affected estimation in stage one models. Underestimation because of inefficiency and

separation due to modelling issues. Overestimation occurred because of overfitting in second stage models for small numbers of trials.

Underestimation worsened as the number of trials increased and sizes decreased. Overestimation largely affected weak surrogacy scenarios which worsened as trial sizes decreased. There was underestimation for small trial sizes and when surrogacy strengths differed at the trial and individual levels and slight differences in results where the ordinal surrogate was not proportional.

Finally, the penalised likelihood technique for dealing with separation was shown to be extremely adept at avoiding bias and loss of information. A mean intercept method used for the penalized technique was adequate for estimation purposes and incurred no bias compared to that used for the trial removal technique.

The fact that the binary-ordinal and ordinal-binary setting results were broadly comparable was not surprising, given that both settings had one dichotomised binary and one categorised ordinal variable.

6.3 Simulation study: conclusions

I ran a simulation study to investigate how well the information theory approach in the ordinal-binary setting estimated surrogacy under a variety of scenarios. Loss of information impacted estimation of surrogacy in the observed ordinal-binary setting at the individual level, with a ceiling of around 0.70. Estimation improved as trial sizes increased and, to a lesser extent, as the number of trials increased. Loss of information in the ordinal-binary setting was not as bad as in the binary-ordinal setting. Seemingly, this is because it is more important to retain information in the surrogate than in the true outcome. Confidence intervals at the individual level were conservative.

At the trial level, overall estimation of surrogacy was good. However, issues occurred in each of the two stages of the approach which impacted results: underestimation, overestimation and separation. Overall, a penalized likelihood technique for dealing with issues of separation in discrete outcomes was far superior to a trial removal technique.

These findings suggest that the information theory approach in the ordinal-binary setting provide a generally effective extension to the surrogacy evaluation framework. The addition of a penalized likelihood technique for dealing with issues of separation is especially worthwhile.

Chapter 7. Extension of information theory to the case of an ordinal surrogate and true outcome: Methodology

In this Chapter I outline how the information theory approach can be extended to the case of an ordinal surrogate and ordinal true outcome (the ordinal-ordinal setting). The extension to the ordinal-ordinal setting incorporates components of the previous two settings. Hence, I will only briefly introduce the information theory approach and how this is extended to the ordinal-ordinal setting. I will refer back to the previous settings to justify the methods adopted.

As in previous chapters I represent a putative surrogate as S, treatment as Z and the true outcome as T. In the multi trial context there are I=1,2,…,N trials, and j=1,2,….,n_i patients per trial. $N_T = \sum_i n_i$ is the total number of patients in all trials. The ordinal surrogate and true outcomes have V and W ordered categories respectively.

Alonso and Molenberghs (2007) suggested the use of the LRF to estimate trial and individual level surrogacy in the multi trial setting under the information theory approach. As in the previous settings I used the LRF to calculate surrogacy.

In the following sections I will remind the reader how the LRF can be used to calculate R_h^2 and R_{ht}^2 in the continuous-continuous setting. I will then show how the LRF can be applied in the case of an ordinal surrogate and ordinal true outcome.

7.1 Individual level surrogacy: LRF reintroduction

The LRF is based on the amount of information gained about the true outcome after accounting for the surrogate. Alonso et al. (2006) proposed two models, for each trial i, for continuous outcomes S and T:

$$T_{ij} = \mu_i + \beta_i Z_{ij} + \varepsilon_{T_{ij}} \quad 7.1$$

$$T_{ij} = \theta_{0_i} + \theta_{1_i} Z_{ij} + \theta_{2_i} S_{ij} + \varepsilon_{T|S_{ij},} \quad 7.2$$

where: θ_{0_i} and μ_i are intercept parameters with and without adjustment for the surrogate; β_i is the treatment effect parameter for the true outcome; θ_{1_i} and θ_{2_i} are treatment and surrogate parameters for the model with adjustment for the surrogate. The difference in the amount of information on the true outcome gained from the surrogate is calculated via the log-likelihood ratio test between 5.1 and 5.2 which is formally expressed as G_i^2, for each trial i. L_0 is always the log-likelihood for the unsaturated model, in this case 5.1, and LL_1 for the saturated model, 5.2, for trial *i*. $G_i^2 = 2 * (LL_1 - LL_0)$. The LRF is then calculated:

$$LRF = R_h^2 = 1 - \frac{1}{N}\sum_i \exp\left(-\frac{G_i^2}{n_i}\right) \quad 7.3$$

R_h^2 can be interpreted as the amount of uncertainty in T that can be explained by knowledge of S, for more details see section 3.3.1.

1.1.1 Individual level: ordinal-ordinal

In the case of an ordinal surrogate and ordinal true outcome, I replaced the models of 5.1 and 5.2 with two proportional odds models, for each trial *i*, for ordinal S and T:

$$logit[P(T_{ij} \leq w)] = \mu_{w_i} + \beta_i Z_{ij} \quad 7.4$$

$$logit[P(T_{ij} \leq w)] = \theta_{w_i} + \theta_{1_i} Z_{ij} + \theta_{2_i} S_{ij} \quad 7.5$$

Where: $w = 1, \ldots, W-1$; *W* is the number of categories in the ordinal true outcome; and μ_{w_i} and θ_{w_i} are the *W-1* intercepts for the cut points of the ordinal true outcome, all other parameters are the same as in the continuous-continuous setting.

As in the ordinal-binary setting, section 5.2.1.1, the ordinal surrogate was modelled as a quantitative explanatory variable.

In the case of discrete true outcomes the information theoretic measure at the individual level was bounded above by a number less than one and had to be rescaled (Alonso and Molenberghs, 2007), this was done in exactly the same manner as in the binary-ordinal setting, see section 3.3.2.

1.2 Trial level surrogacy: LRF reintroduced

In order to apply the LRF at the trial level a two stage approach is required. Again I first describe the case where continuous outcomes are used. At the first stage two models which incorporate all trials are used:

$$S_{ij} = \mu_{S_i} trial_{ij} + \alpha_i (trial_{ij} * Z_{ij}) + \varepsilon_{ij} \quad 7.6$$

$$T_{ij} = \mu_{T_i} trial_{ij} + \beta_i (trial_{ij} * Z_{ij}) + \varepsilon_{ij} \quad 7.7$$

Here: $\hat{\alpha}_i$ and $\hat{\beta}_i$ contain the trial specific treatment effect estimates on the surrogate and true outcome respectively; and μ_{S_i} and μ_{T_i} are the trial specific intercepts for the surrogate and true outcome respectively.

At the second stage two further models are required, the intercept only model, **3.15**, and trial specific treatment effect estimates of true outcome regressed on trial specific intercept and treatment effects of the surrogate, **3.16**:

$$\hat{\beta}_i = \gamma_3 + \varepsilon_i \quad 7.8$$

$$\hat{\beta}_i = \gamma_0 + \gamma_1 \hat{\mu}_{S_i} + \gamma_2 \hat{\alpha}_i + \varepsilon_i, \quad 7.9$$

where: γ_0 and γ_3 are intercept parameters with and without adjustment for the surrogate; γ_1 and γ_2 are the parameters for the surrogate intercept and treatment estimate variables. The difference in the -2*log-likelihood between these two models, G^2, can then be calculated and the LRF applied as in **3.17**.

$$LRF = \widehat{R_{ht}^2} = 1 - \exp\left(-\frac{G^2}{N_T}\right) \quad 7.10$$

1.2.1 Trial level: ordinal-ordinal

In the ordinal-ordinal setting the same procedure for applying the LRF at the trial level is followed as that for continuous outcomes. At the first stage of modelling, since we have an ordinal surrogate and an ordinal true outcome as response variables, proportional odds models are required. These can again be used to calculate treatment effect estimates for each trial.

$$logit[P(S_{ij} \leq v)] = \mu_{S_V}^0 + \mu_{S_i} trial_{ij} + \alpha_i (Z_{ij} * trial_j) \quad 7.11$$

$$logit[P(T_{ij} \leq w)] = \mu_{T_W}^0 + \mu_{T_i} trial_{ij} + \beta_i (Z_{ij} * trial_j) \quad 7.12$$

Where, $v = 1, \dots, V-1$ and $w = 1, \dots, W-1$.

here: $\mu_{S_v}^0$ and $\mu_{T_w}^0$ are assumed fixed intercept cut points relating to the ordinal surrogate and true outcome variable; μ_{S_i} and μ_{T_i} are trial specific shifts of the set of intercepts (Burzykowski et al., 2004); and $\hat{\alpha}_i$ and $\hat{\beta}_i$ are the treatment effect estimates on the surrogate and true outcomes respectively.

At the second stage of modelling, models 3.15 and 3.16 can be fitted and then the LRF, using 3.17, can be calculated in a similar way as for the continuous-continuous case, using μ_{S_i}, $\hat{\alpha}_i$ and $\hat{\beta}_i$. This is exactly the same as in the ordinal-binary setting, see section 3.4.2 for more details.

1.3 Confidence intervals: ordinal-ordinal

Confidence intervals for trial and individual level surrogacy in the ordinal-ordinal setting can be applied in exactly the same manner as the binary-ordinal (and ordinal-binary) setting. See section 3.5.

1.4 Separation: ordinal-ordinal

Separation occurs when there are zero cells in discrete outcome cross tabulations. As in the binary-ordinal and ordinal-binary settings the occurrence of separation could potentially greatly bias trial level results, see section 3.6. A solution to this is to use

the penalized likelihood method of Firth (1993). To see more details of this method see section 3.6.3.

The penalized likelihood method can be applied in the ordinal-ordinal setting in much the same manner as in the ordinal-binary setting (section 5.5). As in the ordinal-binary setting, the calculation of the intercepts is more complicated than in the binary-ordinal setting. A mean intercepts method was adopted in the ordinal-binary setting, see section 5.5, this was found to work adequately and impose no issues of bias. Hence this mean intercept method was used in the ordinal-ordinal setting for the penalized likelihood technique.

1.5 Conclusions: ordinal-ordinal

I have shown the extension of the information theory approach to the ordinal-ordinal scenario at the individual and trial levels and how to apply a penalized likelihood technique to deal with issues of separation at the trial level.

Chapter 8. Simulation study: for an ordinal surrogate and ordinal true outcome

In the previous chapter I discussed how the information theoretic measures R_h^2 and R_{ht}^2 for surrogate evaluation can be extended to the case of an ordinal surrogate and ordinal true outcome. I have conducted a simulation study to investigate how these measures in the ordinal-ordinal setting behave under various scenarios.

In this chapter, I outline the setup of the simulation study and various practical considerations. I then present the simulation results for both R_h^2 and R_{ht}^2.

As in previous chapters: S denotes a surrogate; T a true outcome; Z treatment; with i=1,..,N trials and j=1,..,n_i patients per trial.

8.1 Simulation study: set up

The simulation study for the ordinal-ordinal setting was identical to that for the ordinal-binary setting. Except that both the surrogate and true outcomes were simulated as continuous variables instead of just the surrogate. These were categorised as ordinal outcomes in exactly the same manner as the surrogate in the ordinal-binary setting. Both ordinal variables were simulated as having seven categories. Also, the impact of deviations from the proportional odds assumption needed to be tested at both the individual and trial levels since ordinal response variables were used for all models. The linear relationship (discussed in section 6.1.2.3 in the ordinal-binary setting) and proportional odds assumptions both apply at the individual level. However, this is essentially the same assumption which was tested using the same coding in previous settings, hence for simplicity only the proportional odds assumption is referred to from now on. In all other regards the simulation followed the setup of the ordinal-binary setting. See section 6.1 for more details.

8.2 Simulation study: results

In this section the results of the simulation study for individual level and then trial level surrogacy are presented. The various issues and biases present in the results

were very much in line with those in the binary-ordinal and ordinal-binary settings and have been outlined in Table 8.1.

	Issue	**Scenarios affected**	**Reason**	**See section(s):**
$\boldsymbol{R_h^2}$	R_h^2 estimates substantially lower than true value	All scenarios	Loss of information due to categorisation	4.1.3.3.1 and 6.2.1.5
$\boldsymbol{R_{ht}^2}$	Underestimation	Worse for: • Larger numbers of trials • Strong surrogacy • Small trial sizes	Inefficiency in estimation due to categorisation	4.2.2.1.2 and 6.2.2.1
	Overestimation	Worse for: • Small numbers of trials • Weak surrogacy	Model fitting issue	4.2.2.2 and 6.2.2.2
	Separation	All scenarios	Zero cells in trial crosstabs Resolved through use of Firth (1993) technique	4.1.3.3.4 and 4.2.2.5

Table 8.1: *Issues present in results of simulation study*

I have previously discussed the issues of loss of information, underestimation, overestimation and separation as highlighted in Table 8.1 for the binary-ordinal setting. All issues described in Table 8.1 were present in the results of the ordinal-ordinal setting and are discussed in the sections indicated.

In all the scenarios presented the median R^2 values for the 250 simulations performed for each scenario are given. This is to give some idea of the bias in R^2 estimates. The IQRs of R^2 estimates for the 250 simulations are also given. I present the coverage of and the median lower and upper bounds of the confidence intervals, for each scenario. Results for selected scenarios are presented in this chapter; Appendix C contains full unabridged tables of results for every scenario.

8.2.1 Results: individual level surrogacy

In this section the results for the individual level surrogacy where loss of information impacts results are presented, see section 4.1.3.3.1 and Table 8.1.

Here I present the results for the scenarios:

- where surrogacy was set to be strong at both trial and individual level;
- where this was weak at both levels;
- where the strength of surrogacy differed at trial and individual levels;
- the odds are not proportional.

I also discuss the results of the ceiling investigation for the ordinal-ordinal setting and how the results compare to previous settings.

8.2.1.1 Individual level surrogacy R^2_h: strong surrogacy

Surrogacy was set to be strong at both levels, R^2_h=0.64 and R^2_{ht}=0.90, in the underlying continuous setting. Due to loss of information the observed results for the ordinal-ordinal setting were expected to be lower than in the underlying continuum.

Table 8.2 shows that as the number and size of trial increased the value of surrogacy in the observed ordinal-ordinal setting converged to around 0.53. The impact of trial sizes was greater than that for number of trials, however all values returned were close to 0.53. The IQRs of the R^2_h estimates were much narrower as the number of and size of trials increased. The coverage of the confidence intervals was 100% in all scenarios and the median confidence intervals were very large and covered nearly the whole parameter space where there were ten patients per trial. The precision of these

greatly improved as the size of trials increased, although again the number of trials had a much smaller impact.

Number of trials	Trial size	R_h^2	IQR R_h^2	% cover CIs	Median CI lower	Median CI upper
5	10	0.562	0.175	100%	0.147	0.893
5	60	0.536	0.050	100%	0.322	0.724
5	100	0.536	0.045	100%	0.370	0.687
5	300	0.538	0.035	100%	0.442	0.629
10	10	0.538	0.101	100%	0.133	0.877
10	60	0.534	0.037	100%	0.322	0.722
10	100	0.533	0.033	100%	0.367	0.683
10	300	0.532	0.022	100%	0.436	0.622
20	10	0.539	0.069	100%	0.143	0.880
20	60	0.530	0.029	100%	0.318	0.717
20	100	0.532	0.023	100%	0.366	0.682
20	300	0.530	0.016	100%	0.433	0.620
30	10	0.549	0.055	100%	0.146	0.881
30	60	0.530	0.021	100%	0.318	0.718
30	100	0.528	0.021	100%	0.363	0.679
30	300	0.529	0.013	100%	0.433	0.619

Table 8.2: *Simulation study: Median R_h^2 estimates based on 250 simulations for each scenario, where true values set to: R_h^2 =0.64 and R_{ht}^2 =0.90.*

8.2.1.2 Individual level surrogacy R^2_h: weak surrogacy

The findings for strong surrogacy were mirrored in the case of weak surrogacy, where R_h^2 =0.30 and R_{ht}^2=0.30 in the underlying continuous setting.

Table 8.3 shows that as the number and sizes of trials increased the value of surrogacy in the observed ordinal-ordinal setting converged at around 0.24.

Number of trials	Trial size	R_h^2	IQR R_h^2	% cover CIs	Median CI lower	Median CI upper
5	10	0.293	0.154	100%	0.023	0.736
5	60	0.236	0.051	100%	0.076	0.447
5	100	0.240	0.048	100%	0.104	0.404
5	300	0.241	0.027	100%	0.156	0.336
10	10	0.306	0.100	100%	0.037	0.736
10	60	0.244	0.043	100%	0.080	0.455
10	100	0.245	0.036	100%	0.109	0.408
10	300	0.241	0.019	100%	0.156	0.335
20	10	0.302	0.068	100%	0.039	0.732
20	60	0.243	0.026	100%	0.080	0.453
20	100	0.239	0.023	100%	0.105	0.403
20	300	0.238	0.013	100%	0.153	0.331
30	10	0.308	0.057	100%	0.042	0.734
30	60	0.245	0.022	100%	0.082	0.453
30	100	0.241	0.019	100%	0.106	0.404
30	300	0.239	0.012	100%	0.154	0.333

Table 8.3: *Simulation study: Median R_h^2 estimates based on 250 simulations for each scenario, where true values set to: R_h^2 =0.30 and R_{ht}^2 =0.30.*

8.2.1.3 Individual level surrogacy R^2_h: differing strengths of surrogacy

Table 8.4 shows that R_h^2 estimates for discordant strengths of surrogacy were much the same as those for agreement in surrogacy strengths. The width of the IQRs in the two scenarios was also much the same. The result for weak individual level and strong trial level surrogacy gave comparable results.

Number of trials	Trial size	Surrogacy strong both levels			Surrogacy strong R_h^2, weak R_{ht}^2		
		R_h^2 =0.90	IQR R_h^2	%cover CIs	R_h^2 =0.90	IQR R_h^2	%cover CIs
5	10	0.562	0.175	100%	0.537	0.153	100%
5	60	0.536	0.050	100%	0.534	0.059	100%
5	100	0.536	0.045	100%	0.535	0.042	100%
5	300	0.538	0.035	100%	0.534	0.031	100%
10	10	0.538	0.101	100%	0.538	0.102	100%
10	60	0.534	0.037	100%	0.526	0.038	100%
10	100	0.533	0.033	100%	0.525	0.033	100%
10	300	0.532	0.022	100%	0.528	0.025	100%
20	10	0.539	0.069	100%	0.539	0.076	100%
20	60	0.530	0.029	100%	0.524	0.032	100%
20	100	0.532	0.023	100%	0.525	0.022	100%
20	300	0.530	0.016	100%	0.524	0.017	100%
30	10	0.549	0.055	100%	0.543	0.057	100%
30	60	0.530	0.021	100%	0.526	0.026	100%
30	100	0.528	0.021	100%	0.523	0.020	100%
30	300	0.529	0.013	100%	0.524	0.013	100%

Table 8.4: *Simulation study: Median R_h^2 estimates based on 250 simulations for each scenario, where true values set to: R_h^2 =0.64 and R_{ht}^2 =0.90 or 0.30.*

8.2.1.4 Individual level surrogacy R^2_h: proportional odds assumption

The R_h^2 estimates where the proportional odds assumption was not valid were uniformly slightly lower than those where the assumption holds, regardless of number or size of trials, see Table 8.5.

Number of trials	Trial size	Proportional odds			Non proportional odds		
		R_h^2 =0.90	IQR R_h^2	% cover CIs	R_h^2 =0.90	IQR R_h^2	% cover CIs
5	10	0.562	0.175	100%	0.515	0.155	100%
5	60	0.536	0.050	100%	0.518	0.067	100%
5	100	0.536	0.045	100%	0.526	0.042	100%
5	300	0.538	0.035	100%	0.520	0.032	100%
10	10	0.538	0.101	100%	0.528	0.097	100%
10	60	0.534	0.037	100%	0.514	0.042	100%
10	100	0.533	0.033	100%	0.511	0.038	100%
10	300	0.532	0.022	100%	0.516	0.022	100%
20	10	0.539	0.069	100%	0.535	0.085	100%
20	60	0.530	0.029	100%	0.514	0.028	100%
20	100	0.532	0.023	100%	0.512	0.027	100%
20	300	0.530	0.016	100%	0.510	0.018	100%
30	10	0.549	0.055	100%	0.533	0.057	100%
30	60	0.530	0.021	100%	0.510	0.025	100%
30	100	0.528	0.021	100%	0.510	0.020	100%
30	300	0.529	0.013	100%	0.510	0.013	100%

Table 8.5: *Simulation study: Median R_h^2 estimates based on 250 simulations for each scenario, where true values set to: R_h^2 =0.64 and R_{ht}^2 =0.90. Comparing results for proportional odds and non-proportional odds scenarios.*

8.2.1.5 Individual level surrogacy R^2_h: ceiling effect

The strength of individual level surrogacy in the ordinal-ordinal setting compared the underlying continuum was much lower due to loss of information. I investigated R_h^2 estimated when the strength of surrogacy was 'perfect' at the individual level in the underlying continuous-continuous setting.

As can be seen in Table 8.6 the value of surrogacy of the observed ordinal surrogate converges to around 0.88 as the number and size of trials increased. This suggests that in the case of perfect surrogacy at the underlying continuum the strongest an ordinal surrogate for an ordinal true outcome can be is around 0.88.

Number of trials	Trial size	R_h^2	IQR R_h^2	% cover CIs	lower 95% CI	upper 95% CI
5	10	0.879	0.066	100%	0.495	0.982
5	60	0.888	0.044	100%	0.768	0.946
5	100	0.902	0.038	96%	0.820	0.944
5	300	0.903	0.035	91%	0.862	0.930
10	10	0.864	0.062	100%	0.481	0.977
10	60	0.887	0.036	99%	0.771	0.946
10	100	0.890	0.035	96%	0.806	0.937
10	300	0.892	0.033	85%	0.851	0.923
20	10	0.861	0.033	100%	0.472	0.976
20	60	0.879	0.027	100%	0.761	0.942
20	100	0.885	0.025	100%	0.802	0.935
20	300	0.890	0.028	92%	0.847	0.921
30	10	0.856	0.029	100%	0.467	0.975
30	60	0.878	0.021	100%	0.761	0.942
30	100	0.884	0.024	100%	0.801	0.934
30	300	0.888	0.029	93%	0.846	0.920

Table 8.6: *Simulation study: Median R_h^2 estimates based on 250 simulations for each scenario, where true values set to: R_h^2=1 and R_{ht}^2=0.90.*

As in the binary-ordinal setting, the confidence intervals in the case of 'perfect' surrogacy seems to be behaving slightly differently to other scenarios, see section 4.2.1.5. As in the binary-ordinal setting. the scenario of 'perfect' surrogacy presented here is a very particular case of surrogacy that is unlikely to be seen in practice.

8.2.1.6 Individual level surrogacy R^2_h: comparison to binary-ordinal and ordinal-binary settings

The ceiling effect for the ordinal-ordinal setting was much higher than for the alternative settings when one outcome was binary. In comparison to the underlying continuous-continuous setting the ceiling for the ordinal-ordinal setting was 0.88, the ordinal-binary was 0.70 and the binary-ordinal was 0.48.

Clearly dichotomisation causes a large impact on loss of information compared to categorisation into seven categories. The ceiling in the ordinal-ordinal setting is very

high compared to the other settings where the information in one of the ordinal outcomes has been further reduced to create a binary outcome.

In the context of surrogacy, the amount of information retained in the surrogate was most pertinent to the level of the ceiling. The lowest ceiling witnessed occurred when the surrogate was binary. However, the ceiling in the ordinal-ordinal setting was much higher than the ordinal-binary setting where the true outcome is binary. Therefore, retaining information on the true outcome was also very important.

8.2.1.7 Conclusions individual level surrogacy

When surrogacy strength agrees at both levels R_h^2 estimates converged to around 0.52 where this was set to be 0.64 at the underlying continuum and 0.23 where this was set to 0.30 at the underlying continuum. Smaller trials gave larger R_h^2 estimates, more variability and less precise confidence intervals. The number of trials had limited impact on results. Differing strengths of surrogacy at trial and individual levels and adherence to the proportional odds assumption had little impact on R_h^2 results.

The coverage of the confidence intervals was 100% in all scenarios that were expected to be seen in practice suggesting these are very conservative. Although, the median confidence interval results suggest these were sensible and by no means covered the whole parameter space.

The ceiling effect in the ordinal-ordinal setting was much less pronounced than in other settings. This showed that dichotomisation of continuous outcomes leads to a much larger loss of information than categorisation to ordinal outcomes. In this simulation study the ordinal outcomes had seven categories. However, ordinal outcomes with fewer categories are likely to lead to lower ceilings than those observed in this study.

8.2.2 Results: trial level surrogacy

This section outlines the results for trial level surrogacy using the penalised likelihood technique to deal with issues of separation.

I will describe the scenario where:

- surrogacy was strong at both the trial and individual levels;
- surrogacy was weak at both levels;
- the strength of surrogacy disagred at the trial and individual levels;
- non-proportional odds of the ordinal surrogate and true outcome were present.

Finally, I compare the penalised likelihood technique to the technique where trials with separation were removed from analysis.

8.2.2.1 Trial level surrogacy R^2_{ht}: strong surrogacy

I describe the results for trial level surrogacy when this was strong at both levels and estimates of R^2_{ht} were expected to be in the region of 0.90, see Table 8.7.

Estimation improved as the size of trials increased. Results showed overestimation for five trials but as the number of trials increased estimation improved, especially for large trial sizes. The IQRs of R^2_{ht} and the coverage of confidence intervals improved as the size and number of trials increased. However, the coverage of the confidence intervals was poor where there were large numbers of trials and small numbers of patients per trial, this is likely because of the relative bias in R^2_{ht} results.

Number of trials	Trial size	R^2_{ht}	IQR R^2_{ht}	% cover CIs	Median CI lower	Median CI upper
5	10	0.853	0.228	95%	0.228	0.994
5	60	0.948	0.085	95%	0.507	0.999
5	100	0.959	0.077	92%	0.563	0.999
5	300	0.955	0.079	95%	0.542	0.999
10	10	0.756	0.213	85%	0.274	0.962
10	60	0.892	0.085	99%	0.534	0.989
10	100	0.902	0.087	100%	0.559	0.990
10	300	0.916	0.080	98%	0.597	0.992
20	10	0.682	0.166	48%	0.329	0.897
20	60	0.877	0.070	98%	0.639	0.972
20	100	0.887	0.071	99%	0.659	0.974
20	300	0.899	0.066	99%	0.687	0.978
30	10	0.683	0.114	25%	0.400	0.872
30	60	0.869	0.073	96%	0.680	0.958
30	100	0.880	0.058	98%	0.700	0.963
30	300	0.895	0.047	99%	0.730	0.969

Table 8.7: *Simulation study: Median R^2_{ht} estimates based on 250 simulations for each scenario, in the ordinal-ordinal setting where true values set to: R^2_h =0.64 and R^2_{ht} =0.90.*

8.2.2.1.1 Strong surrogacy: comparison to continuous-continuous setting

In keeping with the results presented in other settings I now compare the ordinal-ordinal results to those for the underlying continuous variables. Table 8.8 shows that the overestimation when the number of trials were small and trial sizes were large was present in both settings. In contrast, the ordinal-ordinal setting showed underestimation in comparison to the continuous-continuous setting when trial sizes decreased. This underestimation worsened slightly as the number of trials increased.

		Ordinal-ordinal setting		Continuous-continuous setting	
Number of trials	Trial size	R^2_{ht}	IQR R^2_{ht}	R^2_{ht}	IQR R^2_{ht}
5	10	0.853	0.228	0.931	0.165
5	60	0.948	0.085	0.950	0.103
5	100	0.959	0.077	0.953	0.095
5	300	0.955	0.079	0.959	0.069
10	10	0.756	0.213	0.840	0.137
10	60	0.892	0.085	0.909	0.096
10	100	0.902	0.087	0.916	0.072
10	300	0.916	0.080	0.919	0.068
20	10	0.682	0.166	0.835	0.100
20	60	0.877	0.070	0.900	0.065
20	100	0.887	0.071	0.902	0.054
20	300	0.899	0.066	0.910	0.058
30	10	0.683	0.114	0.826	0.078
30	60	0.869	0.073	0.895	0.053
30	100	0.880	0.058	0.900	0.050
30	300	0.895	0.047	0.902	0.054

Table 8.8: *Simulation study: Median R^2_{ht} estimates based on 250 simulations for each scenario, in ordinal-ordinal and continuous-continuous setting where true values set to: R^2_h =0.64 and R^2_{ht} =0.90.*

8.2.2.1.2 Strong surrogacy: discussion

As discussed in previous settings the use of a two stage approach at the trial level combined with the loss of information present in categorising continuous outcomes led to inefficiency, see section 4.2.2.1 for more details. Since the estimates converged on the true strength of surrogacy as trial sizes increased this issue was due to lack of efficiency in estimation, rather than an inherent bias.

Further proof of this is provided by the consideration of additional scenarios on top of those mentioned in the methods section. Consider Table 8.9, where 3000 patient scenarios for each trial size have been presented. These results are very similar to

those in the underlying continuum in the 300 patient scenario which indicates that it was indeed inefficiency that caused the underestimation.

		ordinal-ordinal	
Number of trials	Trial size	R^2_{ht}	IQR R^2_{ht}
5	3000	0.958	0.097
10	3000	0.919	0.072
20	3000	0.910	0.052
30	3000	0.906	0.044

Table 8.9 *Additional simulation scenarios: Median R^2_{ht} estimates based on 250 simulations for each scenario, where R^2_h=0.64 and R^2_{ht}=0.90 and trial size was 3000.*

8.2.2.2 Trial level surrogacy R^2_{ht}: weak surrogacy

In this section the case where surrogacy was weak at both levels, $R^2_h = R^2_{ht}$=0.30, is discussed, where I expected the trial level results to be approximately 0.30.

Considering Table 8.10, we see that as trial size decreased overestimation worsened. Where there were five trials and 300 patients' trial level surrogacy was 0.678 which indicates a moderately good surrogate in contrast to the true strength of surrogacy of 0.30. The variation was larger than for strong surrogacy but still showed improvement as the number and size of trials increased. Confidence intervals showed improved coverage as the number of trials increased but were insensitive to increases in trial size.

Number of trials	Trial size	R^2_{ht}	IQR R^2_{ht}	% cover CIs	Median CI lower	Median CI upper
5	10	0.651	0.431	82%	0.030	0.973
5	60	0.609	0.435	82%	0.019	0.967
5	100	0.680	0.442	80%	0.042	0.977
5	300	0.678	0.451	78%	0.041	0.977
10	10	0.372	0.352	94%	0.009	0.819
10	60	0.424	0.353	94%	0.019	0.844
10	100	0.412	0.336	96%	0.016	0.838
10	300	0.452	0.334	95%	0.026	0.857
20	10	0.287	0.255	95%	0.021	0.647
20	60	0.346	0.226	97%	0.045	0.695
20	100	0.340	0.238	96%	0.042	0.691
20	300	0.348	0.241	95%	0.045	0.697
30	10	0.265	0.192	95%	0.038	0.567
30	60	0.328	0.186	97%	0.071	0.623
30	100	0.319	0.186	96%	0.067	0.616
30	300	0.317	0.209	96%	0.065	0.613

Table 8.10: *Simulation study: median R^2_{ht} estimates based on 250 simulations for each scenario, in the ordinal-ordinal setting where true values set to: R^2_h =0.30 and R^2_{ht} =0.30.*

8.2.2.2.1 Weak surrogacy: comparison to underlying continuous-continuous setting

Consider Table 8.11, the results for the ordinal-ordinal setting can be compared to the underlying continuous-continuous setting. Here we see that the pattern of overestimation was consistent at both the observed ordinal-ordinal and underlying continuous-continuous settings. Results in the continuous-continuous setting showed overestimation in every scenario. However, results for the ordinal-ordinal setting were consistently lower than the underlying continuous-continuous setting which, in keeping with the patterns seen for strong surrogacy, worsened as trial sizes decreased.

		Ordinal-ordinal setting		Continuous-continuous setting	
Number of trials	Trial size	R^2_{ht}	IQR R^2_{ht}	R^2_{ht}	IQR R^2_{ht}
5	10	0.651	0.431	0.600	0.453
5	60	0.609	0.435	0.659	0.399
5	100	0.680	0.442	0.676	0.539
5	300	0.678	0.451	0.683	0.440
10	10	0.372	0.352	0.438	0.298
10	60	0.424	0.353	0.446	0.347
10	100	0.412	0.336	0.402	0.376
10	300	0.452	0.334	0.429	0.324
20	10	0.287	0.255	0.349	0.212
20	60	0.346	0.226	0.353	0.199
20	100	0.340	0.238	0.327	0.213
20	300	0.348	0.241	0.355	0.222
30	10	0.265	0.192	0.326	0.172
30	60	0.328	0.186	0.329	0.179
30	100	0.319	0.186	0.358	0.210
30	300	0.317	0.209	0.336	0.204

Table 8.11: *Median R^2_{ht} estimates based on 250 simulations for each scenario, in continuous-continuous setting where true values set to: R^2_h =0.30 and R^2_{ht} =0.30.*

8.2.2.2.2 Overestimation: weak trial level surrogacy

The overestimation witnessed in both the observed ordinal-ordinal and underlying continuous-continuous settings occurred because of overfitting of stage two models for small trial sizes (as in both the binary-ordinal and ordinal-binary settings, see section 4.2.2.2).

8.2.2.2.3 Underestimation: weak trial level surrogacy

There was also evidence of underestimation when surrogacy was weak. Notice that the results in the ordinal-ordinal setting were consistently lower than in the underlying continuous-continuous setting. This occurred because of the inefficiency of a two stage approach with discrete outcomes, this was described in more detail in section 6.2.2.1.

8.2.2.3 Trial level surrogacy R^2_{ht}: differing strengths of surrogacy

Consider Table 8.12, there was very little difference in the results where there was strong surrogacy at the trial level but weak surrogacy at the individual level. The results were lower for the discordant strengths scenario when there were few patients per trial. Confidence intervals coverages were also poorer. Similar discrepant results were found when surrogacy was strong at the individual level and weak at the trial level, see Appendix C Table C.2.

Number of trials	Trial size	Surrogacy strong both levels			Surrogacy strong $\boldsymbol{R^2_{ht}}$, weak $\boldsymbol{R^2_h}$		
		R^2_{ht} =0.90	IQR R^2_{ht}	%cover CIs	R^2_{ht} =0.90	IQR R^2_{ht}	% cover CIs
5	10	0.853	0.228	95%	0.795	0.313	89%
5	60	0.948	0.085	95%	0.930	0.126	96%
5	100	0.959	0.077	92%	0.929	0.119	95%
5	300	0.955	0.079	95%	0.956	0.077	96%
10	10	0.756	0.213	85%	0.634	0.272	63%
10	60	0.892	0.085	99%	0.871	0.119	98%
10	100	0.902	0.087	100%	0.894	0.089	98%
10	300	0.916	0.080	98%	0.906	0.095	99%
20	10	0.682	0.166	48%	0.568	0.218	22%
20	60	0.877	0.070	98%	0.844	0.078	96%
20	100	0.887	0.071	99%	0.874	0.075	99%
20	300	0.899	0.066	99%	0.894	0.069	100%
30	10	0.683	0.114	25%	0.542	0.181	4%
30	60	0.869	0.073	96%	0.833	0.073	92%
30	100	0.880	0.058	98%	0.867	0.071	97%
30	300	0.895	0.047	99%	0.885	0.051	98%

Table 8.12: *Simulation study: Median R^2_{ht} estimates based on 250 simulations for each scenario, where true values set to: R^2_{ht} =0.90 and R^2_h =0.64 or 0.30*

8.2.2.4 Trial level surrogacy R^2_{ht}: non-proportional odds

There were very little differences due to the divergence from the proportional odds assumption on estimates of surrogacy at the trial level, see Table 8.13. R^2_{ht} estimates

were marginally lower when the odds were not proportional. The confidence intervals generally had similar coverages.

Number of trials	Trial size	Proportional			Non proportional		
		R^2_{ht} =0.90	IQR R^2_{ht}	% cover CIs	R^2_{ht} =0.90	IQR R^2_{ht}	% cover CIs
5	10	0.853	0.228	95%	0.850	0.268	90%
5	60	0.948	0.085	95%	0.942	0.127	93%
5	100	0.959	0.077	92%	0.936	0.128	95%
5	300	0.955	0.079	95%	0.943	0.099	94%
10	10	0.756	0.213	85%	0.710	0.180	83%
10	60	0.892	0.085	99%	0.883	0.127	99%
10	100	0.902	0.087	100%	0.891	0.113	99%
10	300	0.916	0.080	98%	0.910	0.082	97%
20	10	0.682	0.166	48%	0.681	0.164	48%
20	60	0.877	0.070	98%	0.860	0.084	98%
20	100	0.887	0.071	99%	0.875	0.074	98%
20	300	0.899	0.066	99%	0.895	0.063	100%
30	10	0.683	0.114	25%	0.673	0.134	24%
30	60	0.869	0.073	96%	0.857	0.060	95%
30	100	0.880	0.058	98%	0.872	0.064	99%
30	300	0.895	0.047	99%	0.881	0.054	97%

Table 8.13: S*imulation study: Median R^2_{ht} estimates based on 250 simulations for each scenario, where true values set to: R^2_{ht} =0.90 and R^2_h =0.64. Comparing results for proportional odds and non-proportional odds.*

8.2.2.5 Trial level surrogacy R^2_{ht}: dealing with separation

8.2.2.5.1 Occurrence of separation: trial level

The percentage of quasi complete separation across scenarios for the ordinal true outcomes in the simulation ranged from 0% to 10.4%, complete separation ranged from 0 to 5.8%. The average was 0% in both cases. In the ordinal surrogate case quasi-complete separation ranged from 0 to 28.8% with an average of 0.5%. Complete separation of S ranged from 0% to 44.5% with an average of 0.7%. The larger occurrences of separation were for small trial sizes.

8.2.2.5.2 Comparison of penalized likelihood and trial removal technique

Number of trials	Trial size	Penalized likelihood technique		Trial removal technique			
		R^2_{ht} =0.90	IQR R^2_{ht}	R^2_{ht} =0.90	IQR R^2_{ht}	% failures	Median trial No.
5	10	0.853	0.228	0.880	0.271	36.4%	4
5	60	0.948	0.085	0.949	0.081	0.0%	5
5	100	0.959	0.077	0.958	0.066	0.0%	5
5	300	0.955	0.079	0.959	0.075	0.0%	5
10	10	0.756	0.213	0.746	0.257	0.4%	7
10	60	0.892	0.085	0.900	0.094	0.0%	10
10	100	0.902	0.087	0.901	0.084	0.0%	10
10	300	0.916	0.080	0.916	0.080	0.0%	10
20	10	0.682	0.166	0.671	0.245	0.0%	14
20	60	0.877	0.070	0.878	0.069	0.0%	20
20	100	0.887	0.071	0.889	0.064	0.0%	20
20	300	0.899	0.066	0.904	0.064	0.0%	20
30	10	0.683	0.114	0.651	0.193	0.0%	30
30	60	0.869	0.073	0.873	0.077	0.0%	30
30	100	0.880	0.058	0.884	0.058	0.0%	30
30	300	0.895	0.047	0.902	0.046	0.0%	30

Table 8.14: *Simulation study: Median R^2_{ht} estimates based on 250 simulations for each scenario, where true values set to: R^2_{ht} =0.90 and R^2_h =0.64. Comparing penalized likelihood technique against trial removal technique (trial removal technique results include the % of time the calculation of R^2_{ht} was not possible and the median number of trials available for analysis when it was)*

The penalized likelihood technique and the trial removal technique (where trials where separation occurred were removed from analysis which led to loss of information) were compared as methods of dealing with separation, see section 3.6.

In Table 8.14 the results for either technique were roughly comparable across settings. However, the variance for the penalized likelihood technique was lower for

small trial sizes where separation occurred most frequently. This was because under the trial removal technique, when separation occurred the trials affected were removed from analysis, resulting in loss of information. Indeed in 36.4% of cases, in the ten patient and five trials scenario, R^2_{ht} could not be calculated at all because fewer than three trials were available for analysis under the trial removal technique. For the scenarios with only 10 patients generally the mean number of trials available for analysis was lower than that set in the simulation. The unavailability of these trials for the trial removal approach resulted in a large amount of information loss which highlights the advantage of the penalized likelihood approach.

The estimation of R^2_{ht} is slightly better for the trial removal technique when trial sizes are large, and separation is likely to be infrequent, with marginally better estimation of surrogacy and lower variance.

8.2.2.6 Trial level surrogacy R^2_{ht}: comparison to binary-ordinal and ordinal-binary settings

In comparison to the binary-ordinal and ordinal-binary settings at the trial level the results for the ordinal-ordinal setting were similar.

However, the underestimation observed in the ordinal-ordinal setting compared to the others was much less severe. In fact, for larger trial sizes the underestimation in the ordinal-ordinal setting was negligible. In the case of 30 trials and 300 patients this was 0.895 when 0.90 was expected, the corresponding result for the binary-ordinal setting was 0.865 and ordinal-binary settings was 0.854. This suggests that a much smaller number of patients per trial gives unbiased estimates for the ordinal-ordinal setting, i.e. estimation in the ordinal-ordinal setting was more efficient.

8.2.3 Results: conclusions

In keeping with the results in previous settings, the ordinal-ordinal setting showed underestimation which worsened as the number of trials increased and trial sizes decreased; there was overestimation especially for weak surrogacy which worsened as trial sizes decreased; and little effect where the proportional odds assumptions was not valid or surrogacy strength differed at trial and individual levels.

Finally, the penalised likelihood technique was more effective at dealing with instances of separation compared to a trial removal approach. The results in the ordinal-ordinal setting showed much reduced underestimation when compared to the binary-ordinal and ordinal-binary settings. This was because both outcomes were ordinal and these were more efficient estimators than the binary outcomes of the other settings.

8.3 Simulation study: conclusions

I ran a simulation to investigate how well the information theory approach in the ordinal-ordinal setting estimated surrogacy under a variety of scenarios.

I found that loss of information impacted estimation of surrogacy in the observed ordinal-ordinal setting at the individual level with a ceiling of around 0.88. Loss of information was reduced compared to other setting since binary outcomes discarded more information than the ordinal outcomes of the ordinal-ordinal setting.

Confidence intervals at the individual level were conservative with 100% coverage for all scenarios which are likely to be seen in practice.

As in previous settings the two stage trial level surrogacy approach suffered modelling issues resulting in: underestimation; overestimation and separation. The penalized likelihood technique worked well at avoiding issues related to separation. Underestimation was much reduced in the ordinal-ordinal setting compared to previous settings since more information is retained in ordinal outcomes.

These findings suggest that the information theory approach in the ordinal-ordinal setting provided a generally effective extension to the surrogacy evaluation framework.

Chapter 9. Case study: All settings

In this chapter I describe the results of a surrogacy investigation I conducted on the stroke clinical trial CLOTS3 (2013) for various forms of discrete true and surrogate variables. This case study was conducted to test the methodology extension of the binary-ordinal, ordinal-binary and ordinal-ordinal settings and to investigate surrogacy in a relevant clinical context.

In this Chapter I:

- introduce stroke;
- introduce the clinical trial CLOTS3 used to investigate surrogacy;
- introduce the primary outcome of interest and potential surrogates for each setting (binary-ordinal, ordinal-binary, and ordinal-ordinal);
- introduce some practical considerations and the clinical and methodological questions of interest;
- discuss how the clinical and methodology questions of interest were addressed following a formal assessment of surrogacy for each setting.

9.1 Surrogacy in stroke

Most strokes are caused by a blockage which cuts off the blood supply to the brain - known as an ischaemic stroke. It is estimated that around four fifths of strokes are ischaemic in white populations (Sandercock et al., 2008). Strokes can also be caused by bleeding in the brain known as a haemorrhagic stroke. In either case loss of blood flow to the brain leads to cell death meaning parts of the brain may stop functioning effectively. This can lead to serious outcomes such as severe disability, permanent vegetative state and death. It has been estimated that stroke causes four million deaths per year worldwide and is the second most common cause of death in developing countries (Murray and Lopez, 1996).

9.2 CLOTS3 trial introduction

I have used the clinical trial CLOTS3 to evaluate potential surrogates for primary outcomes of interest taken at six months in acute stroke patients. In this section I introduce: the terminology; the trial; and its findings.

Venous thromboembolism is the collective term for the occurrence of deep vein thrombosis (DVT), a blood clot in the deep veins of the legs, and pulmonary embolism. Pulmonary embolism (PE) occurs when clots become detached from the veins, travel up through the system and cause blockages to the lungs, which can seriously threaten a patient's life. Venous thromboembolism is an important topic in health research as it is a serious and avoidable complication that causes over 25,000 deaths a year in hospital patients in England (Committee, 2004). This is particularly true of stroke patients, who remain in hospital, since they are generally bedbound and often unable to move one side of their body. After stroke 20-42% of patients suffer a venous thromboembolism (CLOTS, 2013). For this reason, it is particularly important to reduce the occurrence of DVT in stroke patients.

Intermittent pneumatic compression (IPC) aids are inflatable sleeves that are applied to the legs, and are a means of reducing the occurrence of thromboembolism in stroke patients. They compress the legs at fixed time intervals stimulating blood flow around the legs.

CLOTS3 was a 94 centre randomised clinical trial. It was conducted to investigate whether IPC applied to the legs of acute stroke patients reduced the occurrence of DVT (CLOTS, 2013). PE was a secondary outcome of interest since this is not routinely measured and can be difficult to detect. There were 2,876 patients enrolled into the trial and randomised to either IPC or standard care.
CLOTS3 (2013) showed that IPC reduced the odds of DVT by 30 days [OR 0.65 (95% CI 0.51–0.84; p=0.001) after adjustment for baseline variables] and had a positive impact on survival at 6 months, HR 0.86 (0.74–0.99), p=0.042.
I now introduce both a binary and an ordinal primary outcome of interest in stroke clinical trials.

9.3 Case study CLOTS3 set up

In this section I will:

- discuss true outcomes of interest
- introduce potential surrogates
- discuss the clinical surrogacy questions of interest in CLOTS3 for each setting;
- outline the practical considerations required in order to carry out this analyses;
- and discuss two further methodological aspects of information theory in the binary-ordinal setting, illustrated using the case study

9.3.1 Ordinal true outcomes in CLOTS3

A primary measure of treatment effect in stroke patients is the Oxford Handicap Scale (OHS) (Bamford et al., 1989), this is used to assess a patient's ongoing health and survival. The OHS is an ordinal seven point scale of death and disability ranging from no disability to severe handicap and death (see Table 9.1). The OHS is used to assess how well a patient is recovering from stroke. A patient may expect to experience some recovery from their symptoms naturally in the first six months after stroke, any amount of further recovery after a year is less likely. Therefore, a measurement of the extent of disability at six months is common to investigate the likely ongoing disability of a stroke patient. Furthermore, OHS measured at six months has been shown to be a useful indicator of long term survival (Slot et al., 2008).

Clinicians and researchers would find it useful to know which potential surrogates taken at taken early can predict treatment effect on OHS.

Oxford Handicap Scale (OHS)		
None	1	90
Minor symptoms	2	186
Minor handicap	3	312
Moderate handicap	4	626
Moderate to severe handicap	5	366
Severe handicap	6	564
Death	7	697
missing		35

Table 9.1: *Oxford handicap scale: death and disability scale categories*

9.3.2 Binary true outcome in CLOTS3

Another way that one could investigate recovery after stroke is by investigating a binary measure of survival at six months. Clinicians are interested in the causal pathway from stroke to death and investigations of surrogates that lie in this pathway might shed light on these mechanisms of action.

9.3.3 Proposed surrogacy investigation in CLOTS3

As outlined in the previous sections, I wished to investigate potential surrogates for the true outcome, the ordinal OHS measured at six months or the binary survival measure at six months, in acute stroke patients.

In the CLOTS3 trial DVT was one such potential surrogate. As discussed, DVT after stroke can cause death. In less serious cases it can also be a debilitating ailment that may impact a patient's ability to rehabilitate themselves. Therefore, the occurrence of DVT by 30 days as measured in CLOTS3 may serve as some indication of a patient's likely OHS category at six months or their survival.

This surrogate could be formed as either a binary or ordinal measure.

9.3.3.1 Binary surrogate in CLOTS3

In the CLOTS3 trial there were three potential configurations of the proposed binary surrogate DVT in the dataset. These were:

- The occurrence of any DVT by 30 days.

This surrogate captured the causal mechanism of action since it directly measures the occurrence of the event of interest. This surrogate was used as the primary endpoint in CLOTS3 (2013).

- Any DVT or PE by 30 days (DVTPE)

PEs were not routinely collected in this trial and are not easily assessed, hence there are so few recorded see Table 9.2. Hence, the recording of PE was unlikely to provide a great deal of additional information but what was available was still potentially worthwhile.

- Death, any DVT or PE by 30 days (DVTPEDEAD).

• Occurrence of DVT, PE or Death by thirty days	
None	1994
DVT	537
PE	36
Death	345

Table 9.2: *Occurrence of DVT, PE or Death by thirty days*

This surrogate incorporated death without the occurrence of DVT or PE by 30 days. Therefore, these deaths are not related to the causal mechanism of interest (DVT). Therefore, biologically this surrogate was less relevant. However, a measure of death is also incorporated in OHS at six months, therefore, DVTPEDEAD was likely to be more informative than the other surrogates from a statistical perspective. As can be seen in Table 9.2 there are a large number of deaths by thirty days so it seemed likely that it would be important to have them recorded.

9.3.3.2 Ordinal surrogate in CLOTS3

It is possible to consider DVT as an ordinal variable as there are various types of DVTs that can occur and each have different levels of severity.

The occurrence of DVT can be either symptomatic or asymptomatic. Symptomatic DVTs are more serious as they are large enough to be causing symptoms (i.e. pain, swollen limbs etc.). Asymptomatic DVTs can be detected using ultrasound. Two compression duplex ultrasound scans were planned for every patient in CLOTS3. The first ultrasound was conducted between seven and ten days and the second between 25 and 30 days after randomisation. Any occurrence of DVT on either scan was recorded as a DVT by 30 days in the CLOTS3 dataset.

Ordinal DVT surrogate (oDVT)	Value
No DVT	0
Asymptomatic distal (calf) DVT	1
Asymptomatic proximal (Thigh) DVT	2
Symptomatic DVT	3
Pulmonary embolism	4
Death	5

Table 9.3: *Ordinal DVT surogate outcome*

DVTs are also considered more serious if they occur in the thigh as opposed to the calf, referred to respectively as proximal and distal DVTs. PE is a more serious ailment related to DVT that could lead to death. A measure of the severity of these outcomes, as agreed with clinical colleagues, is presented in Table 9.3. This is a six-point ordinal variable of increasing severity of DVT by 30 days (oDVT).

9.3.4 Binary-ordinal setting

In the binary-ordinal setting we assess the three binary measures of DVT occurrence by 30 days for the ordinal OHS assessed at six months.

Ideally a surrogate should encompass both biological and statistical relevancy. Two of the proposed binary surrogates gave biological relevancy, DVT and DVTPE since they represent the biological mechanism of interest. Whereas, DVTPEDEAD was less relevant biologically but perhaps more likely to perform well under formal statistical surrogacy evaluation.

It is hard to know the exact biological workings of stroke recovery. Given this, all suggested binary surrogates were investigated in this setting to see if formal assessment shed some light on the biological mechanisms of action.

Therefore, I performed surrogate evaluation on the three potential surrogates, DVT, DVTPE and DVTPEDEAD measured by 30 days in the binary-ordinal setting.

9.3.5 Ordinal-binary setting

In the ordinal-binary setting an ordinal measure of the severity of DVT, oDVT, by 30 days is used as a surrogate for survival at six months. Occurrence of DVT may cause death therefore in the ordinal-binary setting the surrogate is in the causal pathway to the primary outcome of interest death at six months.

9.3.6 Ordinal-ordinal setting

In the ordinal-ordinal setting the oDVT surrogate is used as a surrogate for assessing OHS at six months.

9.4 Surrogate evaluations overview

In the previous settings I have introduced: CLOTs3; the primary outcomes and surrogates of interest; and the surrogacy investigations proposed for each setting.

In all settings:

- the ordinal OHS or binary survival is the true outcome;
- acute stroke is the clinical area;
- compression aids are the treatment class;
- and a binary or ordinal measure of DVT is the surrogate.

In what follows, I have formally evaluated this surrogate using the information theory approach in the binary-ordinal, ordinal-binary and ordinal-ordinal settings.

9.4.1 Practical considerations

In this section the practicalities of the case study are described.

9.4.1.1 Regrouping centres

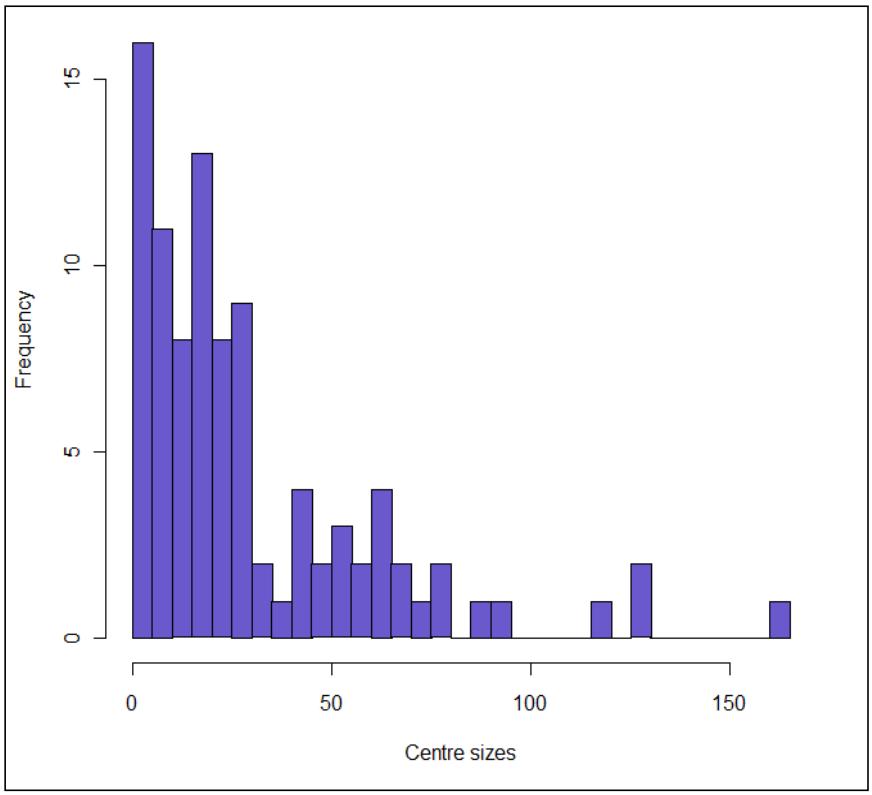


Figure 9.1: *Histogram of centre sizes*

Previously, I discussed surrogacy evaluation in relation to multiple trials but it is equally valid to consider centres within trials (Abrahantes et al., 2004). Therefore, I have based the surrogacy analysis on the centres in the CLOTS3 trial.

CLOTS3 had 94 centres ranging in size from 1-161, see Figure 9.1. There were 25 centres that had very small centre sizes of below ten patients. Twelve of these had fewer than three patients. Models based on centres of very small sizes failed. Furthermore, the smallest trial or centre size considered in the simulation study was ten patients, these scenarios showed poor estimation. Since this was the case, the 25 centres with fewer than ten patients were grouped. They were grouped into four

groups of size: 27, 30, 31 and 32. The median centre size without consideration of the centres with very small centre sizes was 27. The median size after grouping remained 27. Therefore 72 reformatted groups were used in analysis. In what follows these reformatted groups will be referred to as trials to provide consistent terminology in the thesis.

9.4.1.2 Case study: in light of simulation study findings

It was important to make sure that none of the issues of bias reported in the simulation study had an impact on the results of the case study in each setting.

The simulation study showed that underestimation at the trial level increased as the number of trials increased. Since there were 72 trials in the case study underestimation in results could potentially be substantial. As a sensitivity analysis the trials in CLOTS3 were reformatted into 28 groups with a median of 108 patients per group. This is similar to the number and size of groups that were studied in the simulation. The simulation showed that there is a noticeable but minor amount of bias present in this scenario, compared to an unknown amount of bias for a 72 trials scenario.

9.4.1.3 Sensitivity analysis on the causal mechanism of interest

In the binary-ordinal setting three binary surrogates are proposed with and without incorporating death by 30 days, since this is the case we can assess the impact of the inclusion of death by 30 days.

However, in the ordinal-binary and ordinal-ordinal settings the surrogate is a composite since it includes deaths by 30 days as a category and this is also reported in the true outcome (death by six months or OHS). As with the binary-ordinal setting, it is possible that any strength of surrogacy witnessed was driven by the deaths that occurred within 30 days rather than being related to the mechanism of action of interest (DVT). To test this, sensitivity analyses were performed for the ordinal-binary and ordinal-ordinal settings where all patients who died within 30 days were removed from analysis. The sensitivity analyses were applied to this reduced dataset and used to confirm whether DVT severity was the biggest determinant of surrogacy strength in these settings.

9.4.2 Methodological considerations

The two stage nature of R^2_{ht} raises issues such as separation and the failure of the approach to take differences between trials into account, see section 9.4.2.1 and 9.4.2.2. In this section I use the case study to demonstrate the impact of these issues.

9.4.2.1 Separation

In every setting, I was interested in demonstrating why ignoring the occurrence of separation in discrete outcomes leads to bias in trial level surrogacy. Therefore, I will present the results of trial level surrogacy when the occurrence of separation is ignored in the case study, to demonstrate the negative impact it can have. For more information on separation see section 3.6.

9.4.2.2 Weighting

Given that the R^2_{ht} is calculated in two stages the LRF is based only on one G^2, see section 3.4.4.1. This leads to an issue at the trial level since the two stage approach does not adequately take into account the differences between trials. Tibaldi et al. (2003a) suggested adjusting the analysis to account for trial size. This does not totally solve this issue but it takes account of the fact that smaller trials may not calculate treatment effects with enough accuracy. If the treatment effect estimates of smaller trials are considered of equal importance to larger trials in second stage models, more accurate estimates are considered as reliable as inaccurate results and bias could be imposed on results. Weighting allows the estimated treatment effects of larger trials, that are likely to be more accurate, to have greater weight in second stage models – limiting potential bias. It was not necessary to apply this technique to the simulation study as all trial sizes were uniform, however I have done so in the case study as trial sizes vary largely.

In order to adjust for trial size at the second stage of modelling a weighting term can be added to the linear models, see 3.4.2, based on the exact trial size, therefore trials that are larger contribute more to the model. Weighting can be applied to linear models in R using the `weight` term in the function `lm`. The `weight` term tells the model to conduct weighted least squares which in this case minimises $\sum_i weight_i * e^2$ where $weight_i$ is the set weights for each trial i, R Core Team (2016). If $weight_i$ in this case

is set to the exact size of the trial *i* then each trials contribution at the second stage is weighted according to the size of the trial with larger trials contributing more to the analysis.

9.5 Case study investigation aims

The aims of the case study were to:

1. investigate binary and ordinal measures of DVT as surrogates for OHS or survival at six months;
2. demonstrate the impact of ignoring separation in discrete surrogate and true outcomes;
3. demonstrate the impact of weighting by trial size on surrogacy evaluation at the trial level.

9.5.1 Clinical surrogacy investigation

First, I present descriptive statistics of the true and surrogate outcomes in relation to treatment. I then present the formal evaluation of surrogacy for the proposed surrogates using the binary-ordinal, ordinal-binary and ordinal-ordinal information theory approach.

9.5.1.1 Descriptive surrogacy investigation: binary-ordinal

The outcomes of interest in CLOTS3 that are pertinent to our surrogacy investigation are discussed in more detail in the following section. These are the OHS and survival at six months, which are the true outcomes, and the binary DVT surrogates, DVT, DVTPE and DVTPEDEAD at 30 days and the ordinal oDVT.

9.5.1.1.1 Descriptive statistics for primary measures

9.5.1.1.1.1 Treatment by OHS

In general there appeared to be little difference in the OHS when comparing treatment groups as seen in Table 9.4. There was some indication of a slight reduction of the number of deaths, recorded by OHS, for IPC patients and an equal increase in those with severe disability compared to standard care. The reduction in the number of deaths was in keeping with the results of CLOTS3 where a significant improvement in survival was found for IPC patients. Since the treatment benefit only

affected two categories it was possible that the odds would not be proportional (discussed further in section 9.5.1.2).

The number of patients who had missing OHS data were comparable across treatment groups and were removed from analysis, see Table 9.4 and Figure 9.2.

	Standard care		IPC	
	N	Column %	N	Column %
None	45	3.2%	45	3.2%
Minor symptoms	92	6.5%	94	6.6%
Minor handicap	156	11.0%	156	11.0%
Moderate handicap	320	22.5%	306	21.5%
Moderate to severe hand.	185	13.0%	181	12.7%
Severe handicap	255	18.0%	309	21.7%
Death	367	25.8%	330	23.2%
missing	18		17	

Table 9.4: *Oxford Handicap Scale by treatment*

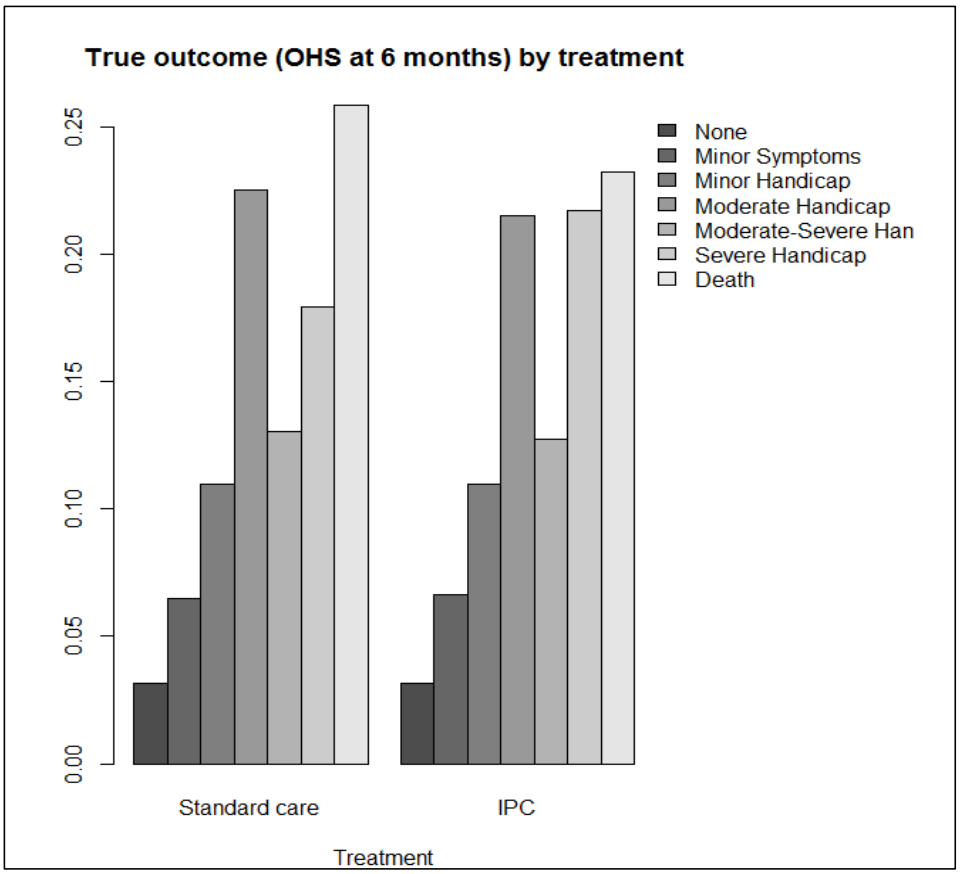


Figure 9.2: O*rdinal OHS true outcome by treatment*

9.5.1.1.1.2 Treatment by survival at six months

Considering the binary true outcome, death by six months, there was a small decrease in the number of patients who died on IPC, see Table 9.5.

	Standard care		IPC	
	N	Column %	N	Column %
No death by 6 months	1078	75.0%	1118	77.7%
Death by 6 months	360	25.0%	320	22.3%

Table 9.5 *Binary true outcome (death by six months) by treatment*

9.5.1.1.2 Descriptive statistics for surrogate measures

9.5.1.1.2.1 Treatment by binary DVT

Here we discuss the three binary DVT measures, DVT, DVTPE and DVTPEDEAD.

There were 537 patients in total who suffered a DVT within 30 days. Only 16.2% of patients on IPC in Table 9.6 suffered a DVT within 30 days as opposed to 21.1% on standard care.

A similar result was seen for DVTPE where 573 patients suffered either a DVT or PE within 30 days, 537 DVT and 36 PE sufferers. Here, 15 patients on IPC had a PE event as opposed to 21 on standard care, see Table 9.6. These results, in light of those for DVT, indicated that IPC reduced the occurrence of DVTPE by 30 days.

Finally, 882 patients suffered DVT, PE or death by 30 days. This means that 309 patients died without suffering a DVT or PE by 30 days. In Table 9.6 we see that 11.5% of standard care patients suffered death by 30 days as opposed to only 9.9% in the IPC group which amounts to 23 fewer events. Again, taking these results in light of those for PE and DVT, this results suggests that IPC reduced the occurrence of DVTPEDEAD by 30 days.

	Standard care		IPC	
	N	Column %	N	Column %
None	947	65.9%	1047	72.8%
DVT	304	21.1%	233	16.2%
PE	21	1.5%	15	1%
DEAD	166	11.5%	143	9.9%

Table 9.6 No adverse events, *DVT, PE and patients who died by 30 days by treatment*

9.5.1.1.2.2 Treatment by oDVT

Considering the surrogate oDVT in relation to treatment it can be seen that there was a reduction in all the oDVT outcomes: DVT; PE; and death by 30 days for IPC patients. There was also an increase in the number of patients who suffer no DVT in comparison to standard care, see Table 9.7 and Figure 9.3.

	Standard care		IPC	
	N	Column %	N	Column %
No DVT	947	65.9%	1047	72.8%
Asympt. distal DVT	92	6.4%	81	5.6%
Asympt. proximal DVT	116	8.1%	79	5.5%
Symptomatic DVT	96	6.7%	73	5.1%
Pulmonary embolism	21	1.5%	15	1.0%
Death	166	11.5%	143	9.9%

Table 9.7: Ordinal *DVT (oDVT) surrogate by treatment*

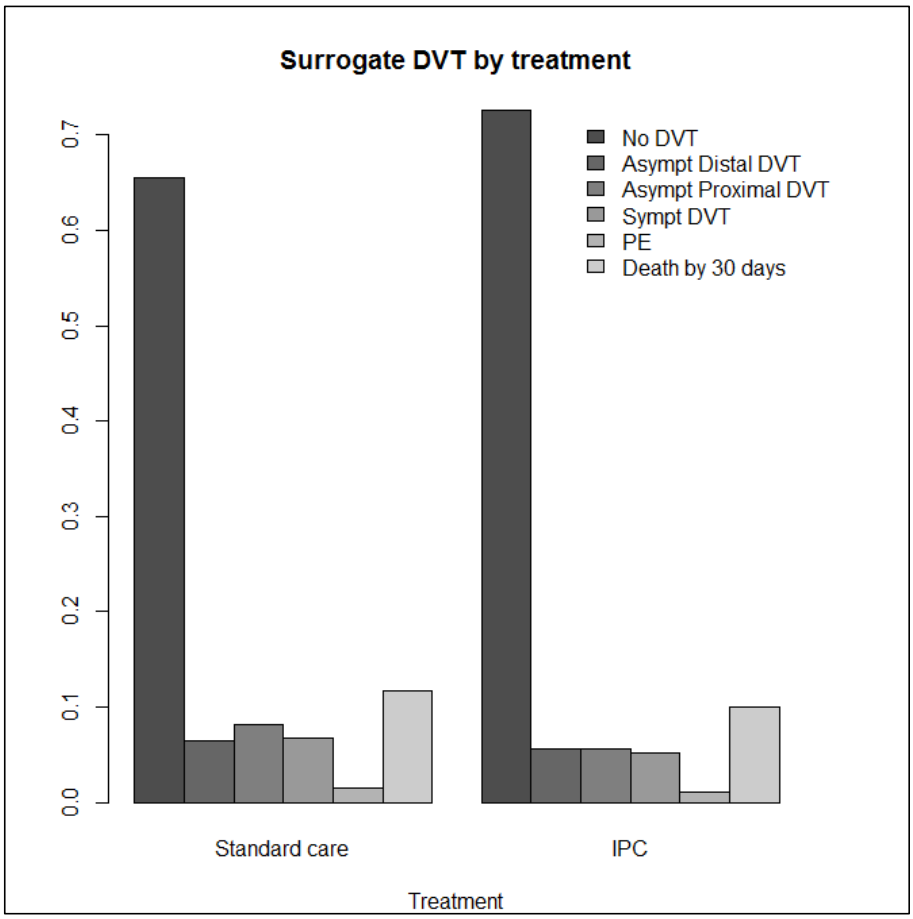


Figure 9.3: oDVT *surrogate by treatment*

9.5.1.2 Descriptive statistics conclusions

All the DVT surrogates both binary and ordinal showed a relationship with treatment. This is in contrast to the results for, true outcome of interest, OHS which did not appear to show a strong relationship with treatment. There was also only a small indication of treatment effect on survival at six months, the other true outcome of interest. Ideally, surrogates should show comparable treatment relationships to the true outcomes of interest.

The descriptive statistics suggested the presence of non-proportional odds for two categories on the seven point ordinal OHS between the treatment groups, see section 9.5.1.1.1.1. The occurrence of deviations of the proportional odds assumption for only two categories on a seven-point scale was specifically investigated in the simulation study. The simulation found that deviations from the proportional odds assumption had very little impact on results. Therefore, even if there were non-

proportional odds in the OHS between treatment groups this should not affect the results of the case study.

9.5.1.3 Formal surrogacy investigation: binary-ordinal

In the following I discuss the results of the formal surrogacy evaluation conducted on the three potential surrogates DVT, DVTPE and DVTPEDEAD for OHS at six months in the binary-ordinal setting.

Surrogate	R_h^2 Individual level	R_{ht}^2 Trial level
DVT	0.049 95% CI (0.001,0.273)	0.077 95% CI (0.003,0.231)
DVT or PE	0.050 95% CI(0.001,0.275)	0.084 95% CI(0.005,0.242)
DVT, PE or death	0.173 95% CI(0.027,0.442)	0.186 95% CI(0.048,0.374)

Table 9.8: Binary-ordinal setting: *Information theory surrogacy results for binary surrogates DVT, DVTPE and DVTPEDEAD*

Table 9.8 gives the formal surrogacy results and Figure 9.4-1.4 shows a graphical representation of the second stage models of information theory at the trial level. At the individual level R_h^2=0.049 and the trial level R_{ht}^2 =0.077 for the binary surrogate DVT. For even a moderate surrogate we would hope to see results over 0.50. The upper limits of the confidence intervals in either case were no higher than 0.30. These results indicate that DVT is not a good surrogate for OHS in stroke patients.

DVTPE had only a few additional surrogate events to those recorded in DVT. As such, the results were very similar. There was also no real difference in the presentation of treatment effects seen in in Figure 9.5. Again, I concluded that DVTPE is not a good surrogate for OHS in stroke clinical trials patients on IPC.

For the surrogate DVTPEDEAD individual level surrogacy was 0.173 and trial level surrogacy was 0.186. The confidence intervals again showed a good degree of precision.

These surrogate results at the trial and individual level were higher than those for both DVT and DVTPE. However, the results were not sufficiently high to suggest that DVTPEDEAD was a good surrogate for OHS at six months in stroke patients.

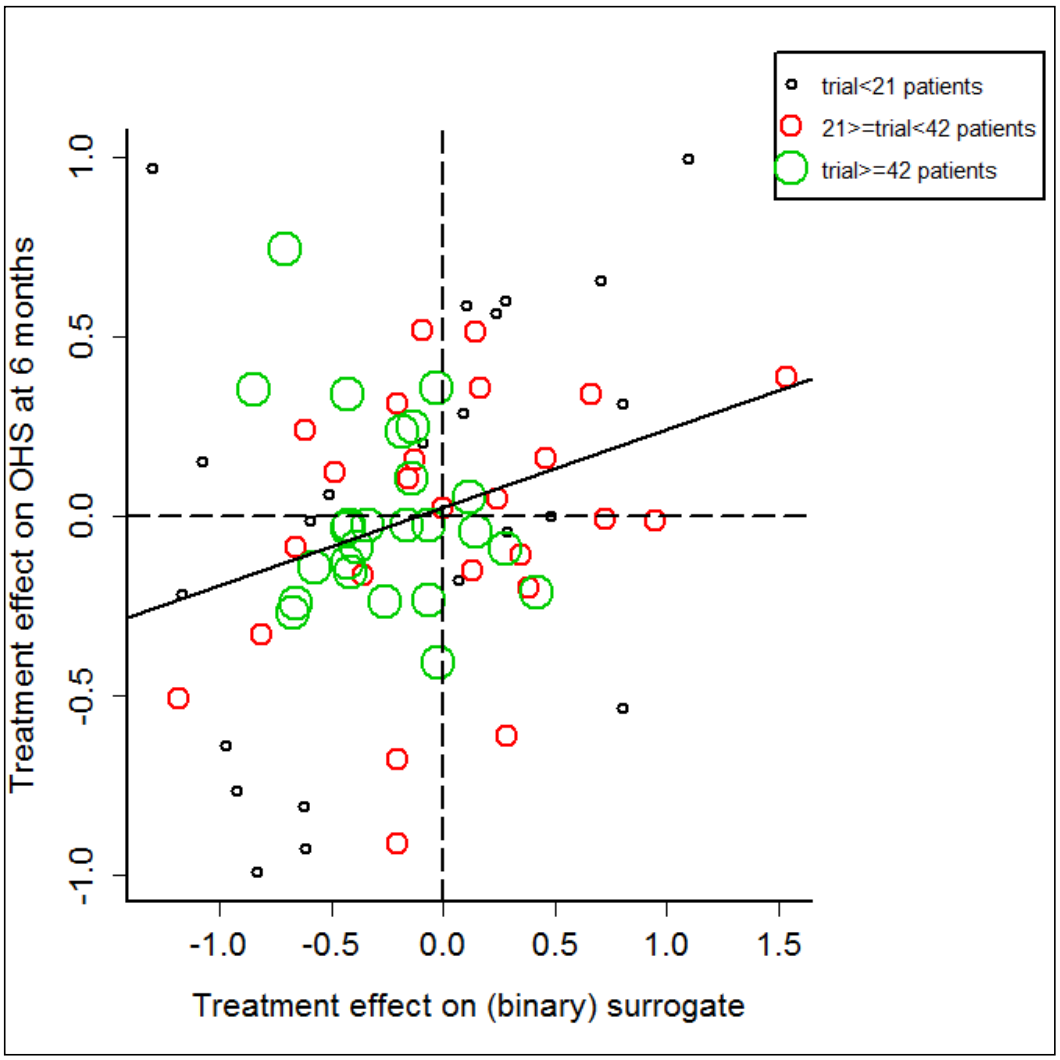


Figure 9.4: Binary-ordinal setting: *Graphical display of trial level surrogacy for binary surrogate indicating patients with DVT; trial size categorisation based on the tertiles of trial size. The regression line represents the regression of the treatment effects of the true outcome on those for the surrogate.*

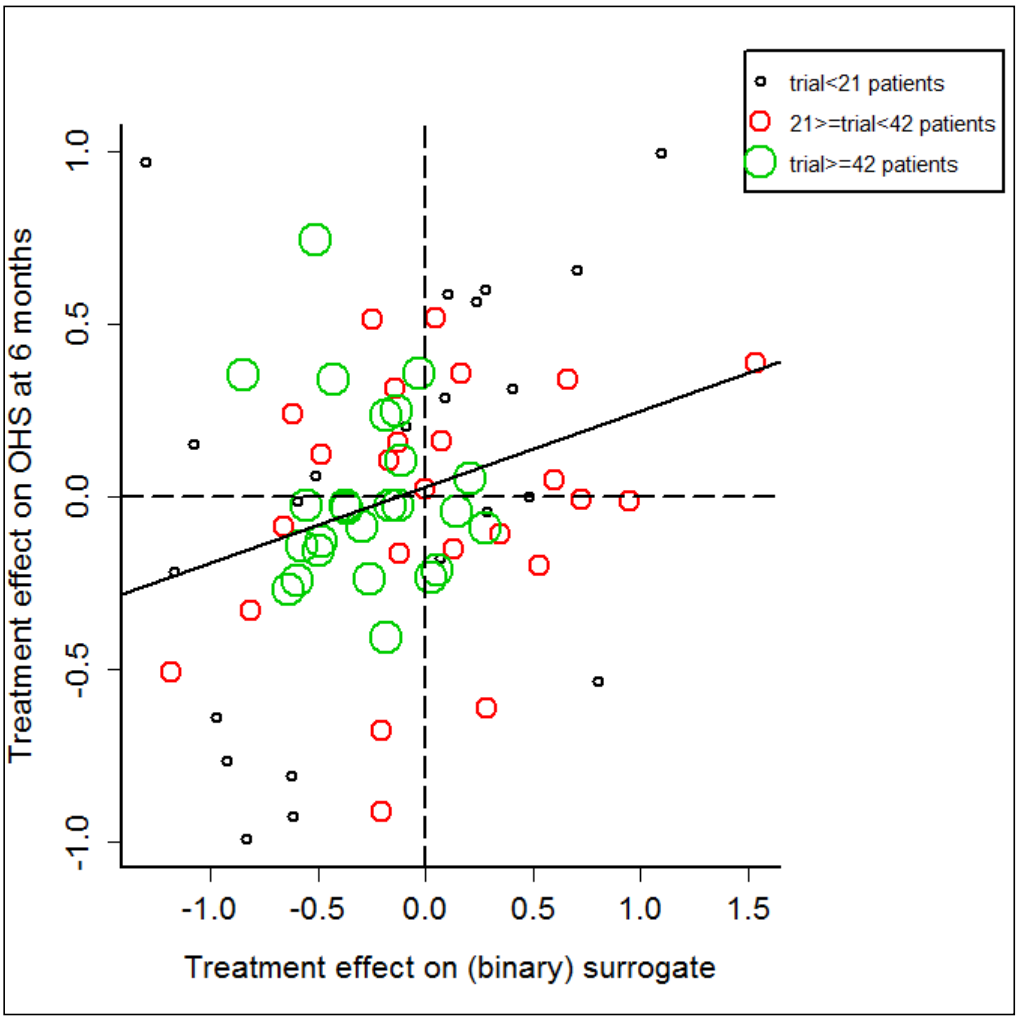


Figure 9.5: Binary-ordinal setting: *Graphical display of trial level surrogacy for binary surrogate indicating patients with DVT or PE; trial size categorisation based on the terciles of trial size. The regression line represents the regression of the treatment effects of the true outcome on those for the surrogate.*

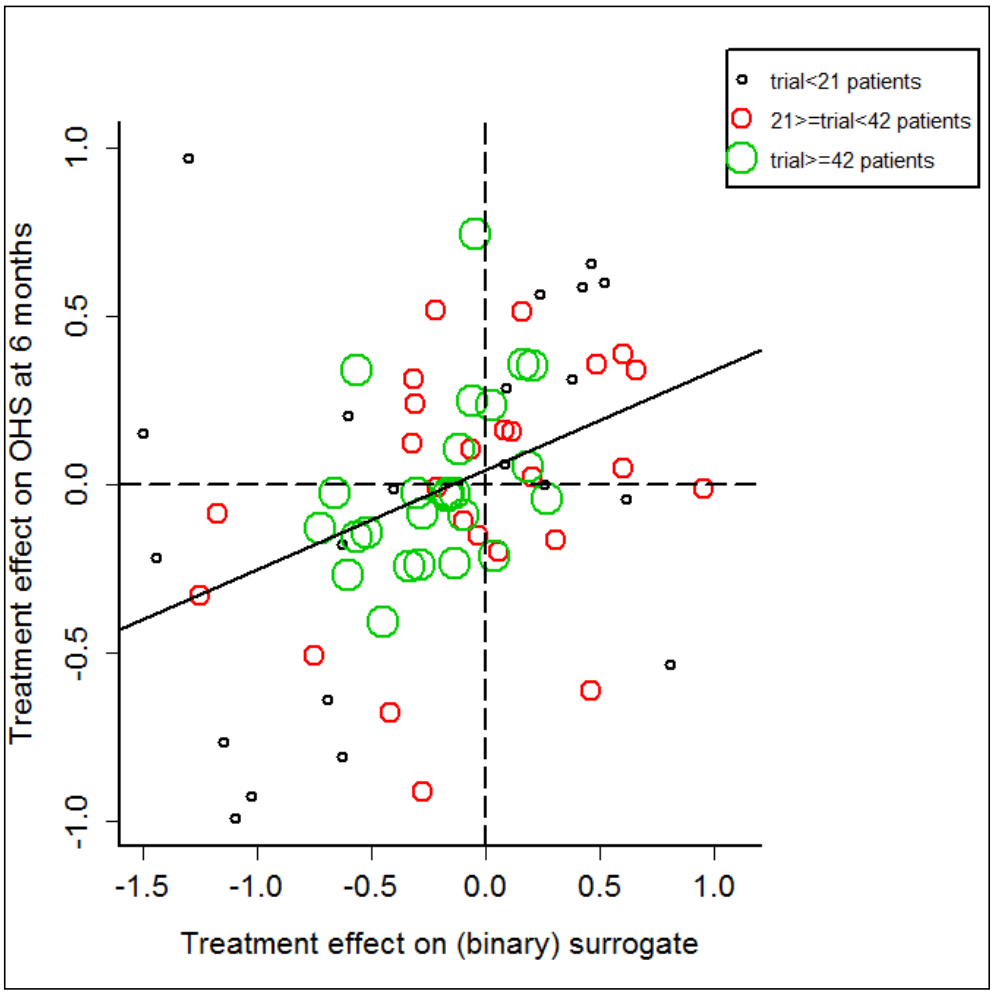


Figure 9.6: Binary-ordinal setting: *Graphical display of trial level surrogacy for patients with DVT, PE or who died by 30 days; trial size categorisation based on the terciles of trial size. The regression line represents the regression of the treatment effects of the true outcome on those for the surrogate.*

9.5.1.4 Formal surrogacy investigation: ordinal-binary

A formal information theory assessment of oDVT, for survival at six months in the ordinal-binary setting, showed that surrogacy was 0.388 at the individual level and 0.315 at the trial level. This can be seen in Table 9.9 and a graphical display of trial level surrogacy in Figure 9.7. Neither result was particularly high. Confidence intervals were fairly wide for individual level surrogacy but these are much narrower for trial level surrogacy.

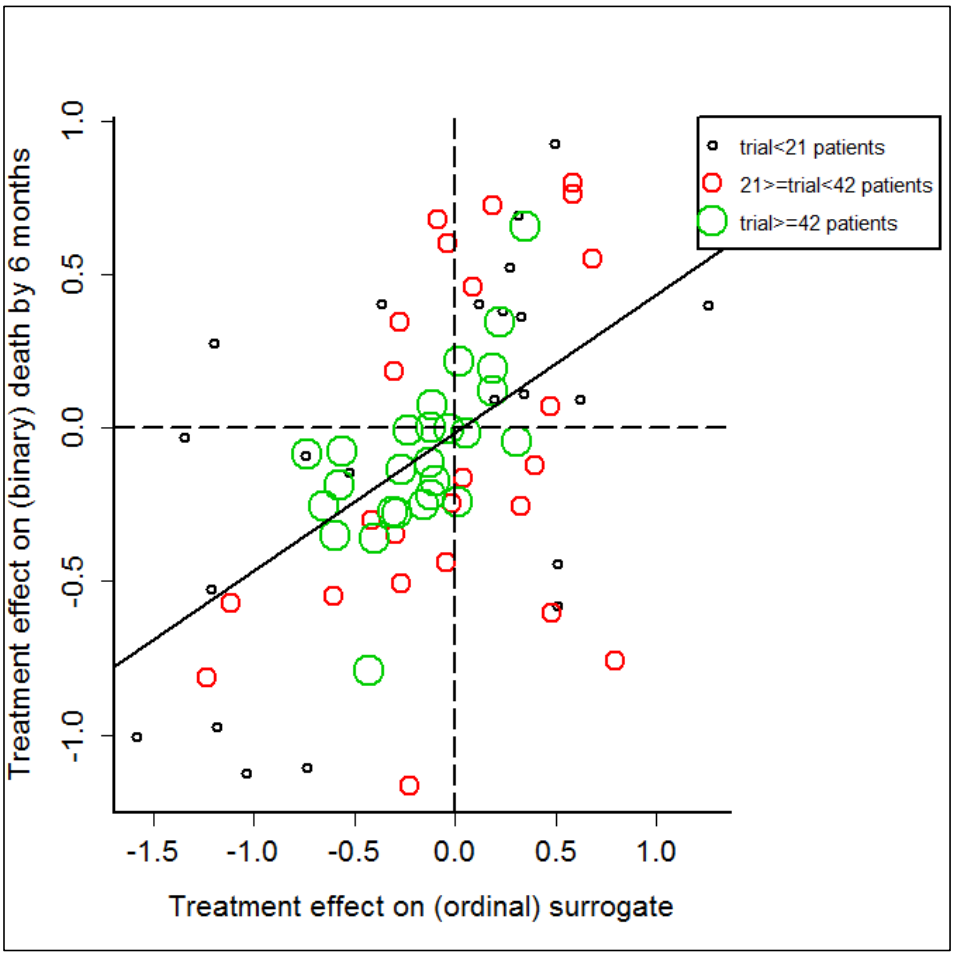


Figure 9.7 : Ordinal-binary setting*: Graphical display of trial level surrogacy for the ordinal DVT surrogate (oDVT); trial size categorisation based on the tertiles of trial size. The regression line represents the regression of the treatment effects of the true outcome on those for the surrogate.*

R_h^2 Individual level	R_{ht}^2 Trial level
0.388 95% CI (0.101,0.812)	0.315 95% CI (0.137,0.511)

Table 9.9: Ordinal-binary setting: *Information theory Surrogacy results for oDVT for binary survival at six months*

9.5.1.5 Formal surrogacy investigation: ordinal-ordinal

The information theory approach for an ordinal DVT as a surrogate for the ordinal OHS showed that surrogacy was 0.257 at the individual level and 0.227 at the trial

level, see Table 9.9 and Figure 9.7. Neither result was particularly high. The confidence intervals suggest that the highest individual level and trial level surrogate values oDVT could take were still only representative of moderate surrogacy.

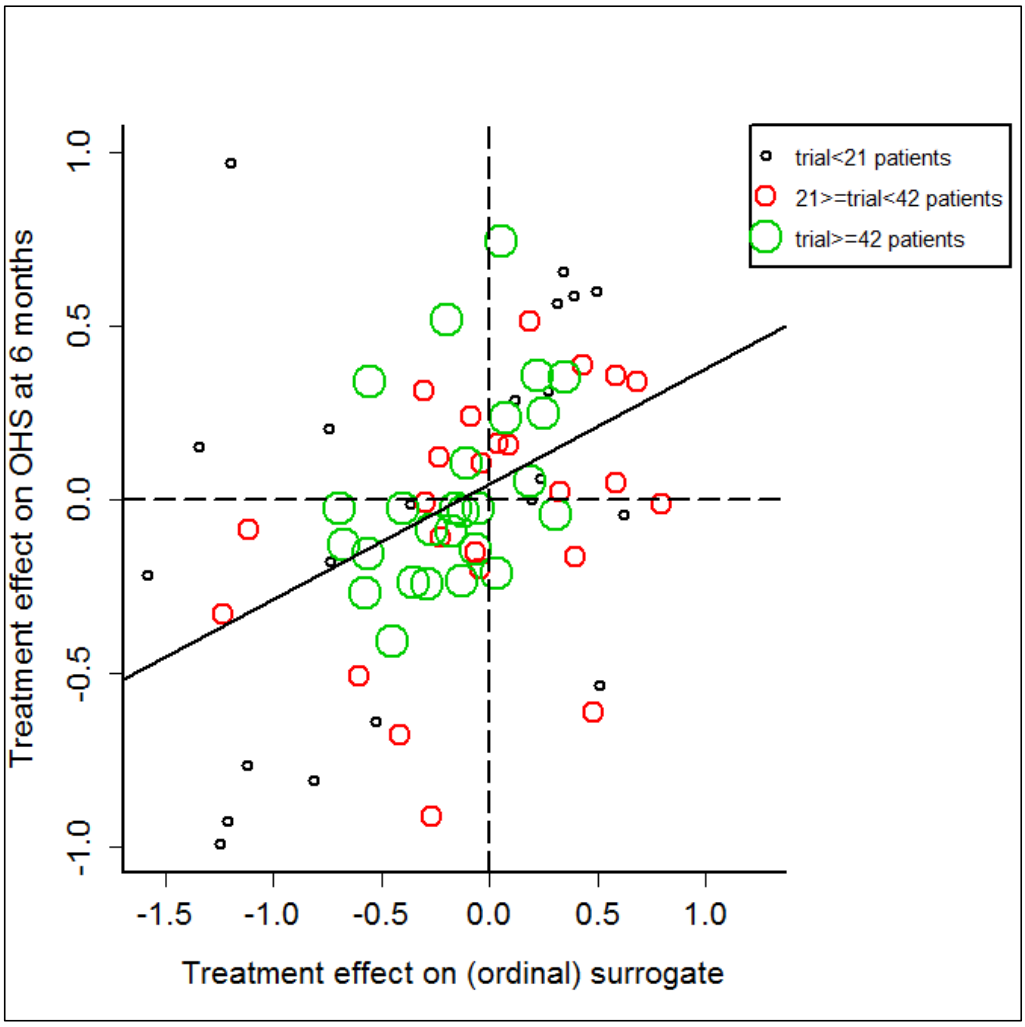


Figure 9.8 : Ordinal-ordinal setting: *Graphical display of trial level surrogacy for ordinal DVT surrogate (oDVT); trial size categorisation based on the tertiles of trial size. The regression line represents the regression of the treatment effects of the true outcome on those for the surrogate.*

R_h^2 Individual level	R_{ht}^2 Trial level
0.257 95% CI (0.062,0.551)	0.227 95% CI (0.074,0.420)

Table 9.10: Ordinal-ordinal setting: *Information theory Surrogacy results for oDVT for OHS*

9.5.1.6 Consideration in relation to simulation study findings: all settings

A sensitivity analysis on a regrouped dataset of 28 groups with a median of 108 patients was conducted to assess whether the underestimation (driven by large numbers of trials) present in the simulation study affected the case study results (see section 9.4.1.2).

	R_h^2 Individual level	R_{ht}^2 Trial level
Binary-ordinal setting		
DVT	0.077 95% CI (0.003,0.231)	0.187 95% CI(0.008,0.494)
DVT or PE	0.084 95% CI(0.005,0.242)	0.296 95% CI(0.048,0.60)
DVT, PE or death	0.186 95% CI(0.048,0.374)	0.351 95% CI(0.079,0.652)
Ordinal-binary setting	0.399 95% CI (0.099,0.430)	0.432 95% CI (0.135,0.716)
Ordinal-ordinal setting	0.247 95% CI (0.107,0.414)	0.327 95% CI (0.065,0.631)

Table 9.11: *Sensitivity regrouping to investigate bias: by setting*

The sensitivity analysis results were very similar to those in the original analyses, see Table 9.11 and Table 9.8 Table 9.10. The trial level surrogacy results were higher in the sensitivity analysis in all settings, suggesting that there was underestimation in

the original analysis. However, the differences were not large enough to change the conclusions in any setting that the proposed surrogates are not good surrogates for the primary outcomes of interest in stoke clinical trials.

9.5.1.7 Sensitivity analysis: ordinal-binary, ordinal-ordinal

A further sensitivity analysis based on concern over the biological relevancy of the proposed surrogate in the ordinal-binary and ordinal-ordinal settings was conducted. To ascertain if DVT was the driving force behind the surrogacy strengths reported, patients that suffered deaths by 30 days were removed from the analysis. See section 9.4.1.3 for more details.

As can be seen in Table 9.12 the surrogacy strength under this sensitivity analysis is very low indeed in either setting. The highest individual level surrogacy was 0.083 and trial level was 0.099 both for the ordinal-binary setting. This suggests that the results presented for the ordinal-binary and ordinal-ordinal settings were driven by the deaths that occur by 30 days rather than the ordinal representation of DVT.

Setting	R_h^2 Individual level	R_{ht}^2 Trial level
Ordinal-binary	0.083 95% CI (0.000,0.133)	0.099 95% CI (0.008,0.268)
Ordinal-ordinal	0.061 95% CI (0.001,0.306)	0.090 95% CI (0.005,0.256)

Table 9.12: *Sensitivity analysis removing deaths: ordinal-binary and ordinal-ordinal settings*

9.5.2 Conclusions on clinical investigation

9.5.2.1 Conclusions: binary-ordinal setting

None of DVT, DVTPE or DVTPEDEAD were strong surrogates for OHS. DVTPEDEAD was the best surrogate statistically because it incorporates patients who died by 30 days and these patients were also a component of the OHS. These deaths do not represent the causal mechanism of action DVT.

Reflecting on the results of the case study in light of the findings of the simulation study did not change the conclusions.

9.5.2.2 Conclusions: ordinal-binary and ordinal-ordinal setting

In the ordinal-binary setting it was found that oDVT by 30 days for the true outcome death by 6 months was a poor surrogate for stroke patients treated by IPC. In the ordinal-ordinal setting, an oDVT surrogate for the true outcome OHS at 6 months has poor surrogate potential for stroke patients treated by IPC.

In both settings sensitivity analyses on regrouped trials showed that the bias that impacted results in the simulation study has limited impact the case study. Further sensitivity analyses indicated that the moderate strength of surrogacy found in both settings was driven by deaths by 30 days which were also present in the true outcome rather than due to the occurrence of DVT. As DVT was the causal mechanism of interest, this suggests that oDVT is an even poorer surrogate in either case than previously concluded.

9.5.2.3 Overall conclusions

In each setting measures of DVT were found to be poor surrogates both proposed primary outcomes of interest. What little surrogacy potential that was found appeared in each setting to be due to the deaths recorded at 30 days that were not part of the mechanism of interest.

Regression diagnostics of trial level surrogacy in second stage models was performed for the main analysis of all proposed surrogates. No major issues with the fit of the models were identified, see Appendix D, D1-5 for plots and discussion.

9.5.3 Methodology considerations: CLOTS3

Below I discuss particular methodological points of interest for trial level surrogacy: separation and weighting by trial size.

In the case of the binary-ordinal setting I use the surrogate results for DVT for demonstrative purposes. Hence there is just one surrogate analyses for each setting.

9.5.3.1 Separation

I present surrogacy results where occurrences of separation were dealt with using the penalized likelihood technique and the converse case where the bias is not dealt with.

Consider Table 9.13, the estimates of R_{ht}^2 - where the penalized likelihood technique was applied – were quite different than where it was not applied. In the binary-ordinal setting the results where the penalised likelihood approach was applied was half that of where it was not. In the ordinal-binary and ordinal-ordinal settings the results were much larger than where the penalized likelihood approach was not applied. These findings suggested that failure to deal with separation leads to erroneous results.

	$\boldsymbol{R_{ht}^2}$ Penalized likelihood	$\boldsymbol{R_{ht}^2}$ No Penalized likelihood
Binary-ordinal setting	0.077 95% CI(0.003,0.231)	0.145 95% CI (0.027,0.325)
Ordinal-binary setting	0.315 95% CI (0.137,0.511)	0.103 95% CI (0.010,0.269)
Ordinal-ordinal setting	0.227 95% CI (0.074,0.420)	0.105 95% CI (0.010,0.273)

Table 9.13: *Results for trial level surrogacy with and without the application of the penalized likelihood technique, by setting*

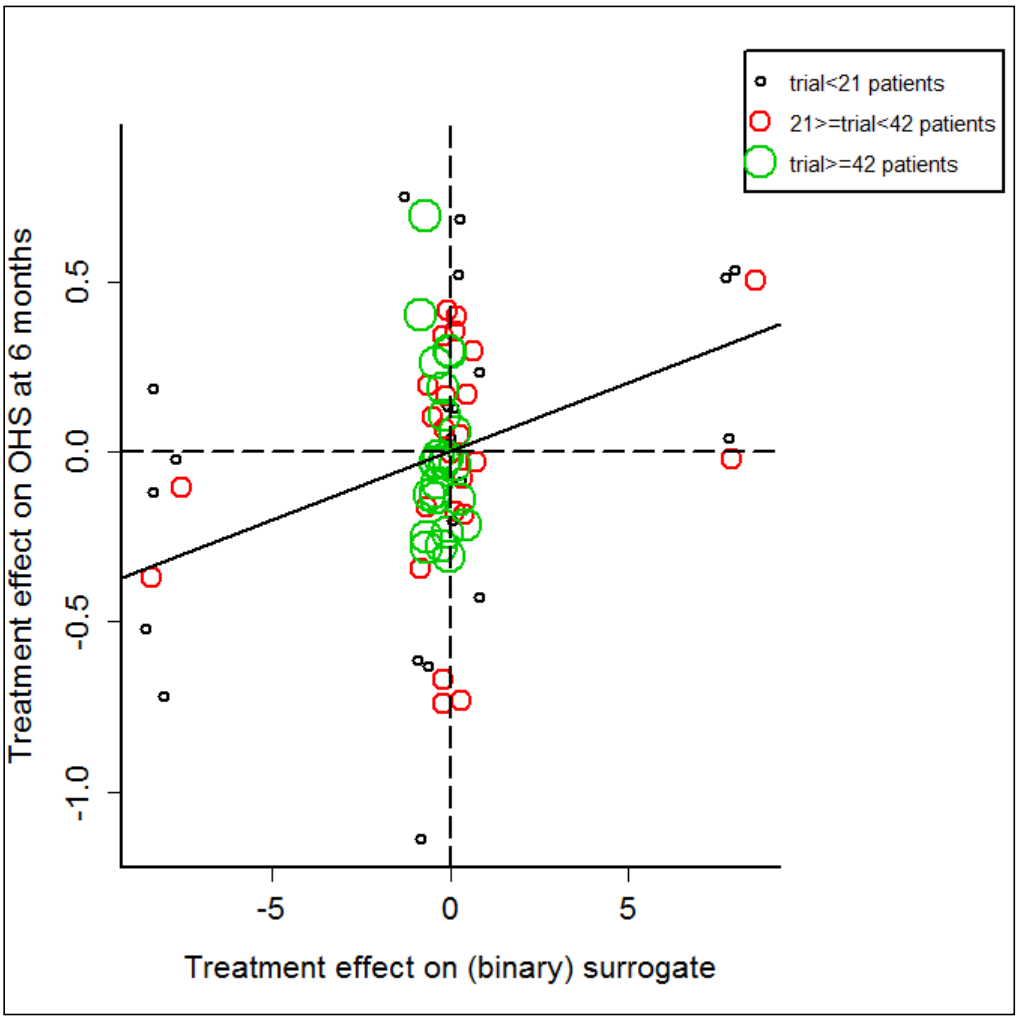


Figure 9.9 *Binary-ordinal setting: Graphical display of trial level surrogacy for DVT where the penalized likelihood technique is not applied; trial size categorisation based on the terciles of trial size. The regression line represents the regression of the treatment effects of the true outcome on those for the surrogate.*

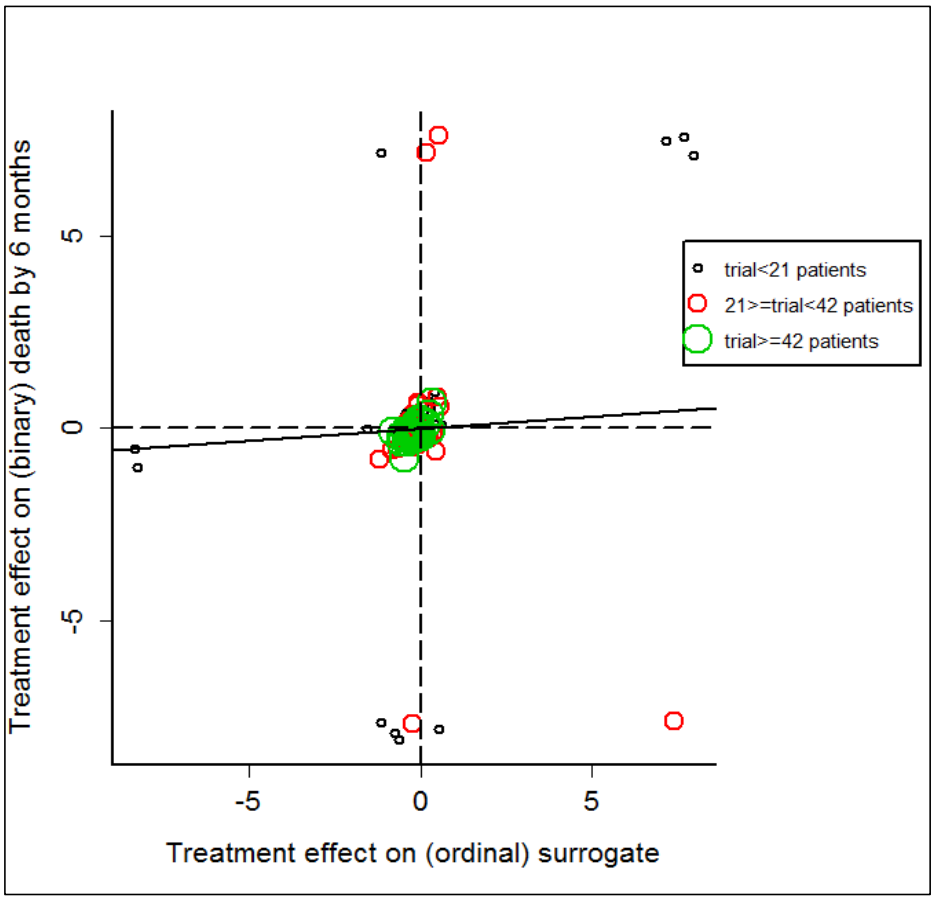


Figure 9.10: Ordinal-binary setting: *Graphical display of trial level surrogacy for ordinal DVT surrogate (oDVT) where the penalized likelihood technique has not been applied; trial size categorisation based on the tertiles of trial size. The regression line represents the regression of the treatment effects of the true outcome on those for the surrogate.*

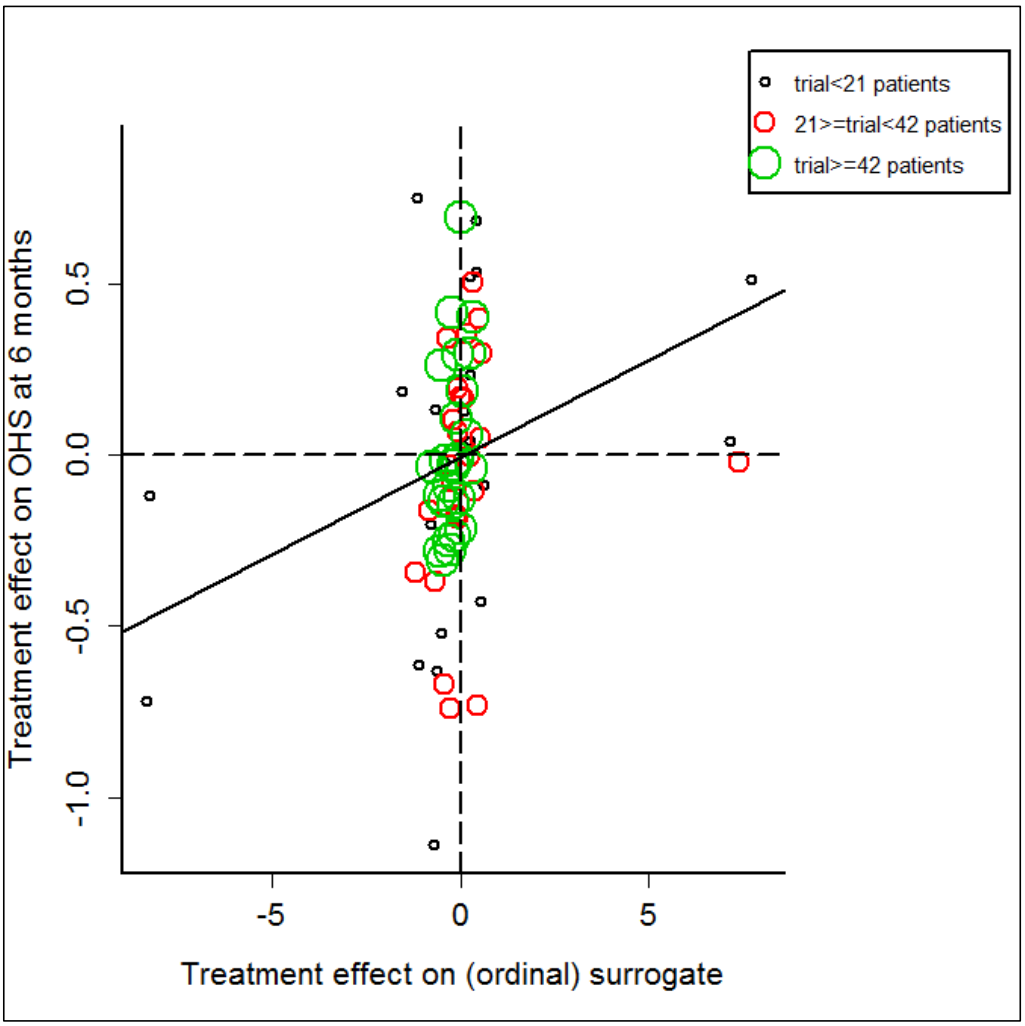


Figure 9.11: Ordinal-ordinal setting: *Graphical display of trial level surrogacy for ordinal DVT surrogate (oDVT) where the penalized likelihood technique has not been applied; trial size categorisation based on the tertiles of trial size. The regression line represents the regression of the treatment effects of the true outcome on those for the surrogate.*

Figure 9.9 -1.9 show the graphical displays of the second stage models of trial level surrogacy in each setting, where separation was not dealt with. Separation occurred: 17 times for the binary surrogate DVT; six times for oDVT; 14 times for the binary true outcome survival at six months; and no times for OHS at six months.

The occurrence of separation at stage one of modelling gives rise to very large, biased estimates of treatment effects (see section 3.6.3). If these estimates are used in

analysis at the second stage they became influential outliers which impact the fit of the model.

This situation can be demonstrated using the case study. In Figure 9.9-1.9 extreme outlying points have clearly influenced the model. The outlying points represent trials where separation occurred and overly large treatment effect estimates were returned. R_{ht}^2 estimates based on models of this form are likely to be biased. Hence, ignoring separation leads to biased surrogacy investigations at the trial level.

Regression diagnostic plots for second stage models for each setting show the impact of these outlying points on the legitimacy of these regressions, see Appendix D, D6-D10.

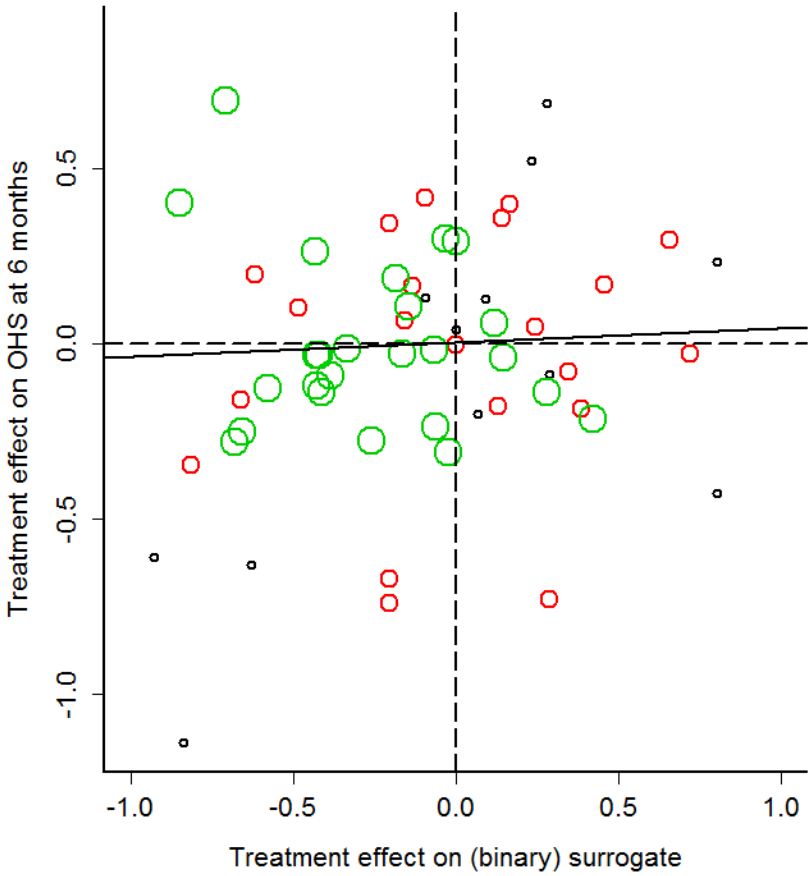


Figure 9.12: *Binary-ordinal setting – x axis scale to match* Figure 9.4*: Graphical display of trial level surrogacy for DVT where the penalized likelihood technique is not applied; trial size categorisation based on the terciles of trial size. The regression line represents the regression of the treatment effects of the true outcome on those for the surrogate.*

To illustrate a comparison of the regressions under the penalised likelihood approach and where separation is ignored we use the DVT surrogacy assessment of the binary-ordinal setting. See Figure 5.10 - where separation was ignored (and the plot put on to the same scale as Figure 5.6)- to Figure 9.4 where the penalized likelihood approach was applied, we can see graphically that the regression line for the treatment effect on the surrogate against those for the true outcome is much different to that in Figure 5.6.The outlying points shown in Figure 9.9 are likely the reason for this.

9.5.3.2 Weighting

A further methodological development that can be demonstrated using the case study data is the weighting of trial level surrogacy by trial size. As stated in section 3.4.4.1 and 9.4.2.2 it is important to take account of trial size in the analysis as smaller trials will have less ability to estimate treatment effects which are used at the second stage of analysis. If the models at the second stage of modelling are unweighted then these smaller trial's treatment estimates will contribute just as much to the model as larger trial's estimates, which are likely to be more precise. Weighting by trial size should enable the model to put more emphasis on estimates that are more reliable.

We can see from Table 9.14 that surrogacy at the trial level for each setting were only slightly different after weighting was applied and confidence intervals overlapped substantially.

	R^2_{ht} Weighting	R^2_{ht} No weighting
Binary-ordinal setting	0.077 95% CI(0.003,0.231)	0.106 95% CI (0.011,0.274)
Ordinal-binary setting	0.315 95% CI (0.137,0.511)	0.297 95% CI (0.123,0.493)
Ordinal-ordinal setting	0.227 95% CI (0.074,0.420)	0.215 95% CI (0.066,0.408)

Table 9.14: *Results for trial level surrogacy with and without weighting by trial size, by setting*

9.5.3.3 Methodological conclusions

I have demonstrated that the occurrence of separation has a serious negative impact on R^2_{ht} estimation. Also, the penalized likelihood approach can effectively deal with this issue.

Weighting by trial size changed the R^2_{ht} results slightly. However, the benefit of weighting may be clearer in an investigation of a wider range of scenarios than was possible in this study.

9.6 Conclusions: case study CLOTS3

None of DVT, DVTPE, DVTPEDEAD or oDVT were worthy surrogates for the true outcome OHS measured at six months or IPC treated acute stroke patients. oDVT was also not a good surrogate for survival at six months.

Methodological issues of interest were effectively demonstrated using the case study.

Chapter 10. Discussion and conclusions

Here I will summarise the results and conclusions of this thesis.

10.1 Motivation

Surrogate outcomes are biological measurements that can be used to predict the true treatment effect on a primary outcome of interest. The use of surrogates can provide shortened, smaller and less invasive trials.

There has been a lot of development in statistical approaches to evaluating surrogates. However, there was previously little methodology for evaluating surrogates where either the surrogate or true outcome were ordinal. This thesis aimed to fill this gap in the literature so that clinical areas such as stroke which routinely use ordinal outcomes (for example the Oxford Handicap Scale) could benefit from the ability to evaluate surrogates.

10.2 Aims

The work of this thesis focused on extending, investigating and refining the current surrogate methodology to the case where outcomes are ordinal in nature. The main aims were as follows:

1. Conduct a systematic review of all the relevant surrogate evaluation methodological literature. This was conducted to determine the best statistical approach to enable evaluation of ordinal outcomes.
2. Extend the selected methodology to situations where either the surrogate, true outcome or both are ordinal.
3. Apply a simulation to assess the resultant surrogate evaluation measures for ordinal outcomes.
4. Illustrate the methodological advancements using a case study in stroke clinical trials data with ordinal outcomes.

10.3 Systematic review conclusions

Through my assessment of the surrogacy literature, and after consideration of the subsequent surrogacy investigations, I conclude, that a good surrogate evaluation approach should:

1. Be practically viable.
2. Be able to inform on the causal nature of treatment effect relationships between the surrogate and true outcome.
3. Identify the surrogate paradox.

   The surrogate paradox occurs when there are positive treatment effects on the surrogate, and a positive relationship between the surrogate and true outcome, but a negative treatment effect on the true outcome, (i.e. the surrogate will validate harmful treatments) see section 2.3.1.3.
4. Inform on the surrogate's transportability or predictive ability.

   Transportable surrogates evaluated in one trial should be able to inform on the treatment effect on the true outcome in a new trial. A surrogate that is unable to do this is useless.

The systematic review identified two main schools of thought in surrogate evaluation methodology. Pragmatic multi-trial approaches (the meta-analytical and information theory approaches) and causality driven approaches (principal stratification and direct and indirect effects approaches).

Baker and Kramer (2003) stated that where outcomes, in the surrogate context, work through multiple pathways (as is often the case) surrogacy assessment is difficult. The drive to identify causal relations between the surrogate and true outcome is therefore a commendable one.

The two main schools of thought deviate in that: causal approaches aim to identify all the causal treatment relationships, and validate surrogates whose causal treatment effects fully agree with the causal treatment effects on the true outcome. At the trial level, the pragmatic multi-trial approaches aim to make sure that the surrogate is able

to predict the treatment effect on the true outcome in a new trial based on the treatment effect on the surrogate. This is regardless of whether the surrogate outcomes fully agree with those of the true outcome. Individual level surrogacy under the multi-trial approach requires agreement in treatment outcomes, though not on a causal basis. Here we see that the causal approaches hold surrogate evaluation up to a much higher standard than the multi-trial approaches. Unfortunately, in practice these higher standards are hard to fulfil.

The current foremost causal approach, principal stratification (PS), requires the calculation of counterfactuals which leads to issues of identifiability. Attempts to resolve the issue of identifiability through the use of strict assumptions can lead to bias. In some very particular circumstances and under strong assumptions this approach is practically viable but this is not true in general. There are also conceptual difficulties with PS. Further checks are needed in addition to original proposals to determine whether surrogates fulfil the requirements of transportability and avoid the surrogate paradox. Practical application of these checks has not been thoroughly investigated and will add more complexity to an already practically difficult approach. Therefore, the causal based approaches only fulfil criterion 2, to a good standard, of the main aims I identified as important for surrogate evaluation approaches.

The multi-trial approaches have a causal interpretation at the trial but not the individual level. This lack of causal rigour is a matter of some concern to authors (Frangakis and Rubin, 2004) and (Joffe and Greene, 2009). However, it could be argued that trial level surrogacy is more likely to be useful to trialists and therefore the lack of causal interpretation at the individual level is not a difficulty (Tibaldi et al., 2003a). Furthermore, the multi-trial approaches are well-established, provide intuitive, informative and practically sound approaches to surrogate evaluation. They can inform on the surrogate paradox unlike the other main approaches (VanderWeele, 2013). Finally, and most crucially, they inform of the transportability of the surrogate which is a vital aim of surrogacy (Joffe, 2011). Therefore the multi-trial approaches fulfil, to a good standard, all of the main criteria I identified as

important for surrogate assessment approaches, unlike the causality based approaches.

Of the two multi-trial approaches the information theory approach is superior to the meta-analytical approach since it provides a consistent interpretation across settings (for example the continuous, binary or ordinal settings). Therefore, this approach is the most worthy candidate for application and for extension to ordinal outcomes.

10.4 Meta-analytical approach: discussion

I will first briefly describe the predecessor to the information theory approach, the meta-analytical approach, which will be useful for context.

The meta-analytical approach was based on a joint mixed model of the surrogate and true outcome regressed on a fixed treatment and a random treatment*trial effect. The joint individual level error variance covariance matrix informed individual level surrogacy and the joint random effects variance covariance matrix informed trial level surrogacy. See section 2.2.1 for models and more information.

There are two issues with this approach:

- First, individual level surrogacy based on this approach was inconsistent across different settings. This issue was the motivation for the development of the information theory approach.
- Second, the joint mixed model came with extreme computational difficulties, which lead to the proposal of a two stage approach. Individual level surrogacy can be calculated at the first stage of this approach but trial level surrogacy can only be assessed at the second stage. This was also the case in the information theory approach which is an analogous approach at the trial level. The reliance on the two stage approach at the trial level in discrete outcomes settings leads to several issues. These issues will be fully discussed in reference to the information theory approach for ordinal outcomes in section 10.6.

The meta-analytical and information theory approaches are intertwined as the latter was set up to approximate the former and the methods are analogous in many

respects. Therefore, a lot of publications for the meta-analytical approach are relevant to the information theory approach.

10.5 Information theory methodology development: discussion

Alonso and Molenberghs (2007) introduced the information theory approach. This uses concepts of information theory to calculate the amount of information gained about the (treatment effects on the) true outcome after adjustment for the (treatment effects on the) surrogate at the (trial) individual level. Under this approach two measures, R_h^2 and R_{ht}^2, estimate surrogacy at the individual and trial levels respectively. These measures fall in the unit interval and ideally both should be close to one to validate surrogacy. Alonso and Molenberghs (2007) took a multi-trial or multi-centre approach and the information theory measures can be calculated via the likelihood reduction factor (LRF) for continuous outcomes. The use of multiple trials or centres within trials would be equally valid for surrogacy evaluation (Abrahantes et al., 2004). In the following, I will refer to trials to represent either case.

10.5.1 Extension of the information theory approach: ordinal outcomes

This methodology has been extended at the individual and trial level to the binary-ordinal, ordinal-binary and ordinal-ordinal settings. This was possible through changing the models from which the LRF is derived in the continuous context to accommodate discrete outcomes.

10.5.1.1 Advancements compared to previous research

The extension of the information theory approach to ordinal outcomes was a novel and useful development of the methodology.

There have previously been two papers on information theory for discrete outcomes one for the binary-continuous (Tilahun et al., 2008b) and the other for the binary-binary setting (Pryseley et al., 2007). This current work has additional advantages to these previous publications as outlined below.

10.5.1.1.1 Individual level

In contrast to the individual level measures of the information theory approach for discrete outcomes in previous research (Tilahun et al., 2008b) and (Pryseley et al., 2007) individual level surrogacy was based on a multi-trial method. This avoided the assumption that the "association between both variables was constant over trials" (Alonso et al., 2004) which may not hold in practice.

10.5.1.1.2 Trial level

If separation occurs in the cross tab of surrogate/true outcome*treatment for a particular trial this can have a serious impact on the calculation of R_{ht}^2. In the presence of separation the treatment effect parameters for the S and T have no unique maximum likelihoods which leads to difficulties in estimation for available software. These programs return extremely large estimates and standard errors. These treatment effect estimates are crucial to the calculation of R_{ht}^2 as they are used at the second stage of modelling as response and explanatory variables. If separation occurs, there are influential outlying points in these second stage models and biased R_{ht}^2 estimates are returned.

After careful consideration of the options I adopted the penalised likelihood technique of Firth (1993) to deal with this issue. Under this technique maximum likelihood estimates of treatment effect estimates are possible and the potential for bias in second stage models is removed. Therefore, trials where separation occurred are able to be retained in analysis and unbiased estimation is possible. This issue was not addressed in previous information theory publications with discrete outcomes (Tilahun et al., 2008b) and (Pryseley et al., 2007).

10.6 Simulation studies: discussion

Simulation studies were conducted for the binary-ordinal, ordinal-binary and ordinal-ordinal settings. This determined how well the approaches might be expected to behave in practice under a range of scenarios. These scenarios:

- varied by the size of trial and the number of trials;
- assessed strong or weak strengths of surrogacy;

- assessed the case of (non) proportional odds
- assessed the case of a (non) linear relationship;
- and represented the case where surrogacy strengths agree or disagree between individual and trial levels.

A thorough investigation was conducted into the best means of setting up the simulation study. One, primarily based on Burzykowski et al. (2005) was adopted based on the joint mixed models of the meta-analysis approach where the surrogate and true outcome were regressed on treatment. The parameters of the model were set to enable simulation and infer certain surrogacy strengths. Continuous surrogates and true outcomes were first simulated under this method which were then dichotomised or categorised to create ordinal or binary outcomes.

10.6.1 Individual level results: discussion

In general, individual level surrogacy performed well in all settings. The size of trial had a bigger impact on results than the number of trials. The IQRs in R_h^2 results were generally narrow indicating good precision in estimation.

There seemed to be little impact on R_h^2 results where the proportional hazards or linear relationship assumptions were invalid. The same was true where surrogacy strengths disagreed across the individual and trial levels.

10.6.1.1 Coverage: individual level

The median confidence intervals showed that the confidence intervals generally had sensible bounds and by no means covered the whole parameter space. However, the coverage of the intervals was 100% in nearly every scenario, that were likely to occur in practice, suggesting that the intervals were conservative.

10.6.1.2 Loss of information: individual level

Under the simulation strategy employed, the individual level results were impacted by loss of information due to the categorisation or dichotomisation of simulated 'underlying' continuous outcomes to 'observed' ordinal or binary surrogate and true

outcomes. This situation was true to life since binary and ordinal outcomes often stand in for an underlying unobservable continuum.

Binary or ordinal surrogates have lower surrogacy potential than their underlying counterparts since they provide less information. It is not possible to calculate the true strength of surrogacy at the individual level for the observed discrete outcomes settings. I added an additional scenario to the simulation in order to investigate the ceilings on the strength of surrogacy possible in the settings involving ordinal and binary outcomes.

I found that discreting continuous variables leads to a great reduction in surrogacy potential. For example, in the worst case of the binary-ordinal setting, surrogacy strength was just less than half as strong as that set at the underlying continuum. A comparison of the impact of loss of information across settings will be given in section 10.6.3.1.

10.6.2 Trial level results: discussion

In general trial level surrogacy estimation performed well. As loss of information does not theoretically affect trial level surrogacy, in discrete settings it should be able to estimate the strength of surrogacy set at the underlying continuum.

10.6.2.1 General results

In general, estimation improved as the size of trials increased. There was little bias imposed by scenarios which deviated from the proportional odds assumption (the linear relationship assumption was not relevant at the trial level). The same was true for scenarios where surrogacy strength differed at the trial and individual levels.

There were three sources of bias found at the trial level. The first of these, separation, has been dealt with by the application of a penalised likelihood approach. The simulation study also identified that both underestimation and overestimation occurred across scenarios. All three of these issues were a result of the two stage estimation of trial level surrogacy and are discussed below.

10.6.2.1.1 Underestimation

Underestimation was worse for small trial sizes and large trial numbers and where surrogacy was strong. Even for large trial sizes of 300 patients the bias was as much as -0.05 for 30 trials.

This underestimation was because of the two stage approach. First, recall that the ability to estimate treatment effects for the true and surrogate outcomes for each trial was crucial in the calculation of trial level surrogacy. Any inefficiency in calculating these treatment effect estimates leads to bias.

The two stage approach of trial level surrogacy can be thought of as splitting up a difficult computational problem into separate parts, solving each part and combining the results. This has been shown to lead to inefficiency in estimation which was worse if the number of partitions was large compared to the size of the partitions (Molenberghs et al., 2011). Partitions in this multi-trial context were trials. The worsening of underestimation with decreases in trial size and increases in number of trials therefore fits with this theory.

Furthermore, the discrete nature of the outcomes in this context compounds this issue as discrete outcomes were shown to be much more inefficient at estimation compared to continuous outcomes (Taylor et al., 2006) and (Taylor and Yu, 2002).

When large sample sizes (3000 patients) for each number of trials scenario were simulated unbiased estimates were returned supporting the theory of inefficiency. Therefore, a combination of the two stage approach and loss of information in discrete outcomes leads to inefficiency which in turn causes underestimation.

10.6.2.1.2 Overestimation

Overestimation was also present and was worse for low strengths of surrogacy and small numbers of trials. In the case of five trials surrogacy can be seriously overestimated which might lead to a false positive validation of a surrogate. This issue was present in results for the continuous-continuous setting also and was therefore not due to the discrete nature of the outcomes investigated.

Second stage models were based on the treatment effect estimates for each trial. Overestimation for small trial numbers was found to be due to models at the second stage suffering overfitting. Instead of modelling the true relationship, the model fits too closely to the data (in this case because there were too few data points). The models appeared to predict the response variable well and return a high R_{ht}^2 even though the true strength of surrogacy was low.

10.6.2.1.3 Separation

The simulation study showed that the occurrence of separation was high in all settings. However, even the occurrence of one instance of separation could seriously bias results if not identified. Therefore, resolving this issue was extremely important.

A penalised likelihood technique was adopted to deal with separation as it allows for unbiased estimation in the presence of separation. It was compared to a trial removal technique and found to be greatly superior: it provided less biased estimation and avoided the large loss of information present in the alternative.

10.6.3 Comparison across settings: discussion

The results as discussed above were applicable to all settings investigated in this thesis. However, there were particular differences in the results of the binary-ordinal, ordinal-binary, and the ordinal-ordinal settings.

10.6.3.1 Differences across settings: individual level

At the individual level the main difference between settings was in the ceiling effect. The ceiling for the ordinal-ordinal setting was around 0.88, the ordinal-binary was around 0.70 and the binary-ordinal was around 0.48 when surrogacy is set to be perfect, R_h^2=1, in the underlying continuum.

Cox (1957) shows that the maximum amount of information retained when a dichotomisation occurs was 63.7%. The retention of information increased for categorical outcomes as the number of categorises increased. The reported largest amount of information retained was 94.2% for six categories: our ordinal variable has seven categories. These findings agree with my results as more information was

retained for the ordinal-ordinal setting where both outcomes were ordinal. The ceiling in the ordinal-ordinal setting is around 0.88 which shows a good amount of retention of information.

In comparison the binary-ordinal and ordinal-binary setting have much lower individual level surrogacy ceilings. We see that the amount of information retained in the surrogate was most pertinent to the height of the ceiling. The lowest ceiling witnessed occurred when the surrogate was binary in the binary-ordinal setting. This is around 0.48 which demonstrates a very large decrease in surrogacy potential for the binary surrogate. Which suggests that at the individual level binary surrogates cannot provide enough information to act as good surrogates.

Retaining information on the true outcome was also very important as the ceiling in the ordinal-ordinal setting was much higher than where the true outcome is binary in the ordinal-binary setting.

In summary, loss of information in the surrogate or true outcome has a negative impact on surrogacy strength at the individual level. Ordinal outcomes on a seven point scale were vastly preferable to binary outcomes which do not provide very good information on surrogacy at the individual level.

10.6.3.2 Differences across settings: trial level

At the trial level results generally differed only in terms of underestimation and separation. Results for overestimation were much the same.

10.6.3.2.1 Differences: underestimation

Underestimation was not as severe for the ordinal-ordinal setting.

As previously mentioned, underestimation was partially driven by inefficiency due to the loss of information in the discrete outcomes compared to the underlying continuum.

Where regression variables were categorised inefficiency occurs, this was much worse for binary outcomes compared to ordinal outcomes since they suffer greater loss of information (Taylor et al., 2006) and (Taylor and Yu, 2002). Since, ordinal

outcomes were not as inefficient at estimation the underestimation was not as bad for the ordinal-ordinal setting.

10.6.3.2.2 Differences: separation

Separation occurred less frequently for the ordinal outcomes in the simulation. Hence, instances of separation were less in the ordinal-ordinal setting where both outcomes were ordinal.

10.6.4 Comparison of simulation to previous research: discussion

The simulation studies of this thesis incorporated a more thorough investigation of scenarios than previous simulation studies for information theory approaches. In addition to scenarios investigated by previous authors, I included scenarios for: weak strengths of surrogacy; assessment of differing strengths of surrogacy at the trial and individual levels; the investigation of ceiling values for individual level surrogacy; and extremely large trial sizes to investigate inefficiency.

10.6.4.1 Previous research: individual level

In this research a multi-trial surrogate evaluation method at the individual level was adopted. This was not used in previous discrete surrogate research for the binary-binary or binary-continuous settings (Tilahun et al., 2008a), (Pryseley et al., 2007) respectively, see section 4.2.1.6.

R_h^2				
Binary-binary	Binary-ordinal	Binary-continuous	Ordinal-binary	Ordinal-ordinal
0.215	0.293	0.16	0.389	0.529
(Tilahun et al., 2008a)	(Present study)	(Pryseley et al., 2007)	(Present study)	(Present study)

Table 10.1: *Median R_h^2 results for 30 trials and 300 patients where the true strength of surrogacy was R_h^2=0.64. Decimal places were to three dp except for binary-continuous case where this was not provided.*

In Table 10.1 we see the information loss for strong surrogacy across previous studies and in my research for the intuitive best scenario of 30 patients and 300 patients. The settings in this table was ordered, from left to right, according to the increasing amounts of information expected to be retained in each (expectations were driven by the results of my work where retention of information in S as opposed to T had the bigger impact).

Results at the individual level show much more loss of information in the binary-continuous setting, R_{ht}^2=0.16 (Pryseley et al., 2007), than any of the settings I investigated. This was lower than expected as the continuous true outcome does not suffer loss of information therefore the results of this setting should, at least, be higher than the binary-binary and binary-ordinal settings. Results for the binary-binary setting (Tilahun et al., 2008a) were much lower than the results in all the settings I have investigated. However, this might be explained by the fact that more information was retained in the outcomes I investigated compared to the two binary outcomes of the binary-binary setting.

However, the evidence of more impact of loss of information in the binary-binary and binary-continuous settings compared to those of this thesis may be explained, at least in part, by the fact that they have not adopted a multi-trial method. A multi-trial surrogate evaluation method avoids the strong assumption of constant association

across trials and is more in keeping with the ethos of the information theory multi-trial approach.

10.6.4.2 Comparison to previous research: trial level

10.6.4.2.1 Previous research: underestimation

Underestimation was witnessed in previous research however it was not possible to compare this to my research since the vast majority of this was likely to be due to unresolved occurrences of separation. The impact of separation will be discussed in section 10.6.4.2.3.

10.6.4.2.2 Overestimation

In the meta-analytical approach, Burzykowski et al. (2005) investigated moderate surrogacy strength R^2_{ht}=0.50 for continuous outcomes. Overestimation of around 0.04 was witnessed for the 25 trials and 50 patient scenario under simulation. This was a similar amount of overestimation as seen in my results for similar scenarios. At the trial level I would expect patterns of behaviour to be very similar between the meta-analytical and information theory approaches. These results therefore corroborate my findings to some extent.

10.6.4.2.3 Separation

In the case of surrogacy evaluation for discrete outcomes neither the binary-binary or binary-continuous settings (Tilahun et al., 2008a) and (Pryseley et al., 2007) respectively, dealt with the issue of separation. As such the results at the trial level suffered extreme underestimation in comparison to the ordinal settings I investigated.

In the case of 30 trials and 300 patients per trial for high strengths of surrogacy a median R^2_{ht} of 0.764 was given in the binary-binary and 0.75 in the binary-continuous setting when 0.90 was expected (Tilahun et al., 2008a) and (Pryseley et al., 2007) respectively. This was in comparison to 0.865 in the binary-ordinal, 0.854 in the ordinal-binary and 0.895 in the ordinal-ordinal setting results for the same scenarios in my work.

These previous publications, therefore, demonstrate the serious negative impact of ignoring the presence of separation in evaluating surrogacy.

10.7 Simulation study: conclusions

The simulation study provided a more thorough examination of a more comprehensive range of scenarios than previous research based on the information theory approach.

In general, trial and individual level surrogacy evaluation performed well using the information theory methodology for ordinal outcomes developed in Chapter 3, Chapter 5 and Chapter 7.

Loss of information meant that individual level surrogacy for the observed discrete outcome setting was negatively impacted. Ordinal outcomes were less impacted than binary outcomes especially in regard to the surrogate outcome. Potentially a multi-trial approach led to improved results compared to previous literature (Tilahun et al., 2008a) and (Pryseley et al., 2007) although this may be due to other factors. Confidence intervals for individual level surrogacy where found to be conservative.

The two stage nature of trial level surrogacy led to three issues: overestimation; underestimation; and separation. Overestimation was due to overfitting in second stage models. Underestimation and separation were due to issues in estimation in stage one models.

Of these three issues, separation had the potential to cause the most serious bias. If separation occurred even once within a trial and went unnoticed the R_{ht}^2 results were likely to be unsound. A penalised likelihood approach was applied to address these issues which removed bias and avoided the serious issues of loss of information. In comparison to ignoring the problem as was done in previous literature (Tilahun et al., 2008a) and (Pryseley et al., 2007) bias was reduced. This was a highly beneficial development for surrogacy evaluation in discrete settings.

10.8 Case study: discussion

I used the CLOTS3 (2013) trial to investigate ordinal surrogates and true outcomes using the information theory methodology under all the theoretical settings investigated in this thesis.

DVT is a serious event which causes pain and discomfort and may lead to life threatening events. A routine measurement of outcome for stoke patients is the Oxford Handicap Scale (OHS) typically measured at six months. The OHS is a measure of death and disability and records the severity of ongoing amounts of disability of patients who suffered a stroke. I was interested in whether some measurement of DVT may be a surrogate for death at six months or OHS.

CLOTS3 (2013) investigated whether intermittent pneumatic compression aids (IPC) applied to the legs reduced the occurrence of DVT by thirty days. I used this trial to investigate whether:

- a binary indicator of occurrence of DVT by 30 days was a surrogate for OHS at six months
    - using the information theory approach for the binary-ordinal setting
- an ordinal measure of DVT by 30 days was a good surrogate for death at six months
    - using the information theory approach for the ordinal-binary setting
- an ordinal measure of DVT by 30 days was a good surrogate for OHS at six months
    - using the information theory approach for the ordinal-ordinal setting.

10.8.1 Clinical findings: discussion

I found that none of the proposed surrogates were good surrogates for OHS at six months for stroke patients. What little surrogacy potential was available was mostly driven by deaths within 30 days that were recorded as part of the DVT surrogate measures, rather than driven by the mechanism of interest DVT. Therefore, DVT was not a good surrogate in its own right for either death or OHS at six months in stroke clinical trials patients on IPC.

10.8.2 Methodological findings: discussion

10.8.2.1 Separation

The case study was used to demonstrate the benefits of the penalized likelihood approach to deal with instances of separation. The graphical displays of trial level surrogacy where separation was not dealt with showed the extreme impact on surrogacy evaluation imposed by this issue.

10.8.2.2 Weighting

Tibaldi et al. (2003b) suggested using a weighting technique, based on trial size, at the second stage of trial level surrogacy to partially account for differences between trials. I provided results with and without weighting and found limited impact on surrogacy estimation.

10.9 Case study: conclusions

I found that both binary and ordinal measures of DVT were not good surrogates for death or OHS at six months in stroke clinical trials patients on IPC.

The case study showed that separation frequently occurred in real life scenarios. The results also told a cautionary tale of the dangers of ignoring instances of separation when evaluating surrogacy for discrete outcomes. It also showed how effectively the penalized likelihood removed the instances of bias associated with separation.

The usefulness of weighting by trial size to partially account for differences between trials was not strongly demonstrated in this work.

10.10 Future work

This thesis has provided a useful extension of surrogate evaluation using the information theory approach and a thorough examination of the various ways in which this works in practice. It has identified areas where: investigation to give further insight into the mechanics of this approach are needed; and development of the methodology to overcome issues raised in this work is required.

10.10.1 Confidence intervals: individual level

The confidence intervals at the individual level were found to be conservative. Bootstrap intervals were suggested by Tilahun et.al. (2008a) to give improved precision in confidence intervals for the information theory approach in the binary-binary setting. Therefore, the development of bootstrap intervals in the ordinal settings covered in this thesis may resolve the issue of the conservative confidence intervals seen in my work.

10.10.2 Number of ordinal categories

In the simulation study, I showed that the use of seven point ordinal outcomes gives superior results to binary outcomes in a number of different ways: first, loss of information was not as severe at the individual level; second, underestimation was not as bad at the trial level; and finally, separation does not occur as frequently.

However, a seven point ordinal outcome retains a lot of information compared to ordinal outcomes with less categories. A full investigation of ordinal outcomes with a range of numbers of categories would better demonstrate how evaluation of ordinal surrogates and true outcomes would behave in real life scenarios.

10.10.3 Trial level

I have demonstrated that the two stage approach at the trial level has led to three issues with trial level surrogacy: separation; underestimation and overestimation. Previous research used random effects models at stage one or a weighting method at stage two to partially deal with potential issues in the two stage approach (Tibaldi et al., 2003b). Another alternative, not previously proposed, is to consider a Bayesian analysis. These three potential solutions are now considered in turn.

10.10.3.1 Trial level: weighting

Tibaldi et al. (2003b) suggested using a weighting technique based on trial size, see section 10.8.2.2, to partially account for differences between trials in the calculation of R_{ht}^2.

The simulations used in this thesis adopted the approach of previous research and used trials of equal sizes in each scenario. Therefore, it was unnecessary to apply a weighting technique based on trial size.

In the case studies I have applied the weighting technique so that differences in trials were partially accounted for. The benefit of this approach was not strongly demonstrated. However, these applications were only under the conditions of one specific scenario where surrogacy was generally weak and the true strength of surrogacy was unknown. Therefore, in order to determine the true benefit of the weighting technique it might be useful to conduct a simulation study for various scenarios where the size of trial varies within each simulated set of trials and the true value of surrogacy is known.

10.10.3.2 Trial level: random effects

Previous approaches have used mixed models at the first stage of the two stage approach to potentially provide better estimation of treatment effects, suggested in Tibaldi et al. (2003b). These were applied for discrete outcomes in the information theory approach in Pryseley et al.(2007) and Tilahun et al. (2008b). Such an approach might limit the underestimation witnessed in my results if a random effects approach is more efficient at estimation.

We could hypothetically have adopted a random effects approach for our ordinal outcome settings. I have done some preliminary investigation into the feasibility of this. Where there were ordinal outcomes, under current statistical software, it was possible to model random effects for proportional odds models. However, it was only possible to get an overall random treatment effects estimate for trial*treatment. Unfortunately, the information theory approach requires treatment effect estimates for each trial. In the frequentist context it is difficult to integrate over random effects for discrete responses (McCulloch and Searle, 2001). Approximations of the estimates needed are possible for binary outcomes but the present available software would suggest this is not the case for ordinal outcomes. Therefore, these estimates cannot be provided. There were packages available that provided these results using a Bayesian approach but initial investigation returned nonsensical results for

uninformative priors. In order to make a random effects approach possible for ordinal outcomes more investigation of the Bayesian option is needed.

A Bayesian approach could be a big undertaking for potentially little reward (i.e. if a random effects approach has little additional benefit in estimation compared to the fixed effects approach).

10.10.3.3 Trial level: Bayesian joint mixed model

The main difficulties of discrete outcome surrogacy evaluation at the trial level were due to the two stage nature of the information theory approach. The two stage approach was necessary because the original joint modelling approach of the meta-analytical approach was too computationally burdensome. In the original proposal trial level surrogacy was based on the variance covariance matrix of a frequentist joint mixed model, see section 10.4. Therefore, only a one stage approach was needed. If it had been computationally effective this original proposal would have been the ideal means of evaluating surrogacy at the trial level. Under this approach the issues encountered at stage two of the approach I used would not occur.

Potentially, a Bayesian one stage meta-analytical approach could resolve the computational issues of the frequentist approach as Bayesian analysis is more flexible than frequentist methods. Developing and testing a Bayesian method would potentially be a large undertaking and may also have a high computational burden.

10.11 Conclusions

In conclusion, this work provides an important investigation of the methodological approaches for evaluation of surrogacy which carefully examined the advantages and disadvantages of the main approaches. It also provides an essential extension of the means of evaluating surrogacy for ordinal outcomes.

The information theory approach, for ordinal outcomes, has been extremely well investigated via simulation and case study. The simulation studies were much more thorough than any previous publications in information theory in terms of the scenarios investigated and the attention given to understanding patterns in results (the

latter point proved by the oversight of previous authors (Tilahun et al., 2008b) and (Pryseley et al., 2007) to identify the presence of separation or highlight the bias in results).

These investigations have identified the probable ceiling strengths of binary and ordinal surrogates at the individual level and three previously unknown issues relating to the calculation of trial level surrogacy. These issues are: separation and underestimation which were particular to discrete outcome settings; and overestimation which was a serious issue present in all settings. Of the three the most extreme bias was imposed by separation; the occurrence of which can have an extremely detrimental impact on results. The separation issue has been resolved through the application of a penalized likelihood approach, which performed extremely well in the simulation and case studies. The other issues of overestimation and underestimation were not so easily remedied. Random effects and weighting approaches may provide partial improvement for the underestimation but a joint mixed model Bayesian approach may be the best solution for both these issues.

To summarise: the methodological work of this thesis has: filled a gap in the literature to provide clinicians, trialists and researchers with the means of evaluating surrogates in the case of ordinal outcomes; been shown to work well in practice; and provided a solution to the serious issue of separation. The simulation study investigations will help researchers to better understand the behaviour of the information theory approach under a wide range of scenarios. It has also identified issues in estimation that require resolution. Finally, the case studies conducted ruled out DVT as a potential surrogate for death or OHS at six months in stroke patients which will be of interest to clinicians working in this area.

References

ABRAHANTES, J. C. & BURZYKOWSKI, T. 2010. Simplified modeling strategies for surrogate validation with multivariate failure-time data. *Computational statistics & data analysis,* 54**,** 1457-1466.

ABRAHANTES, J. C., MOLENBERGHS, G., BURZYKOWSKI, T., SHKEDY, Z., ABAD, A. A. & RENARD, D. 2004. Choice of units of analysis and modeling strategies in multilevel hierarchical models. *Computational statistics & data analysis,* 47**,** 537-563.

ABRAHANTES, J. C., SHKEDY, Z. & MOLENBERGHS, G. 2008. Alternative methods to evaluate trial level surrogacy. *Clinical Trials,* 5**,** 194-208.

AGRESTI, A. 2002. Categorical Data Analysis (Wiley Series in Probability and Statistics).

AGRESTI, A. 2012. *Software supplement for categorical data analysis* [Online]. Available: http://www.stat.ufl.edu/~aa/cda/software.html [Accessed 14/12/2014 2014].

AGRESTI, A. 2014. *Categorical data analysis*, John Wiley & Sons.

ALLISON, P. D. 2008. Convergence failures in logistic regression. *In:* SAS Global Forum, 2008. Citeseer, 1-11.

ALONSO, A., GEYS, H., MOLENBERGHS, G., KENWARD, M. G. & VANGENEUGDEN, T. 2004a. Validation of surrogate markers in multiple randomized clinical trials with repeated measurements: canonical correlation approach. *Biometrics,* 60**,** 845-853.

ALONSO, A., GEYS, H., MOLENBERGHS, G. & VANGENEUGDEN, T. 2002. Investigating the criterion validity of psychiatric symptom scales using surrogate marker validation methodology. *Journal of biopharmaceutical statistics,* 12**,** 161-178.

ALONSO, A. & MOLENBERGHS, G. 2006. Surrogate marker evaluation from an information theory perspective. *Biometrics,* 63**,** 180-186.

ALONSO, A. & MOLENBERGHS, G. 2008. Evaluating time to cancer recurrence as a surrogate marker for survival from an information theory perspective. *Statistical methods in medical research,* 17**,** 497-504.

ALONSO, A., MOLENBERGHS, G., BURZYKOWSKI, I., RENARD, D., GEYS, H., SHKEDY, Z., TIBALDI, F., ABRAHANTES, J. C. & BUYSE, M. 2007. Quantifying the effect of the surrogate marker by information gain - Reply. *Biometrics,* 63**,** 962-963.

ALONSO, A., MOLENBERGHS, G., BURZYKOWSKI, T., RENARD, D., GEYS, H., SHKEDY, Z., TIBALDI, F., ABRAHANTES, J. C. & BUYSE, M. 2004b. Prentice's Approach and the Meta-Analytic Paradigm: A Reflection on the Role of Statistics in the Evaluation of Surrogate Endpoints. *Biometrics,* 60**,** 724-728.

ALONSO, A., MOLENBERGHS, G., GEYS, H., BUYSE, M. & VANGENEUGDEN, T. 2006. A unifying approach for surrogate marker validation based on Prentice's criteria. *Statistics in medicine,* 25**,** 205-221.

ALONSO, A., VAN DER ELST, W., MOLENBERGHS, G., BUYSE, M. & BURZYKOWSKI, T. 2014. On the relationship between the causal-inference

and meta-analytic paradigms for the validation of surrogate endpoints. *Biometrics, 71 (1), 15-24.*

BAKER, S. G. 2006. A simple meta-analytic approach for using a binary surrogate endpoint to predict the effect of intervention on true endpoint. *Biostatistics,* 7**,** 58-70.

BAKER, S. G. 2008. Two simple approaches for validating a binary surrogate endpoint using data from multiple trials. *Statistical methods in medical research,* 17**,** 505-514.

BAKER, S. G. 2009. Two simple approaches for validating a binary surrogate endpoint using data from multiple trials (vol 17, pg 505, 2008). *Statistical methods in medical research,* 18**,** 227-227.

BAKER, S. G., IZMIRLIAN, G. & KIPNIS, V. 2005. Resolving paradoxes involving surrogate end points. *Journal of the Royal Statistical Society Series A-Statistics in Society,* 168**,** 753-762.

BAKER, S. G. & KRAMER, B. S. 2003. A perfect correlate does not a surrogate make. *BMC Medical Research Methodology,* 3**,** 16.

BAKER, S. G. & KRAMER, B. S. 2012. Surrogate Endpoint Analysis: An Exercise in Extrapolation. *Journal of the National Cancer Institute,* 105 (5), 316-320.

BAKER, S. G., SARGENT, D. J., BUYSE, M. & BURZYKOWSKI, T. 2012. Predicting treatment effect from surrogate endpoints and historical trials: an extrapolation involving probabilities of a binary outcome or survival to a specific time. *Biometrics,* 68**,** 248-57.

BAMFORD, J., SANDERCOCK, P., WARLOW, C. & SLATTERY, J. 1989. Interobserver agreement for the assessment of handicap in stroke patients. *Stroke,* 20**,** 828-828.

BEGG, C. B. & LEUNG, D. H. 2000. On the use of surrogate end points in randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 163**,** 15-28.

BENDA, N. & GERLINGER, C. 2007. Sperm count as a surrogate endpoint for male fertility control. *Statistics in medicine,* 26**,** 4905-4913.

BERGER, V. W., IZMIRLIAN, G. & KNOLL, D. 2012. Discussions. *Biometrics,* 68**,** 239-241.

BERGMAN, J. & HOLMQUIST, B. 2012. A measure of dependence between two compositions. *Australian & New Zealand Journal of Statistics,* 54**,** 451-461.

BOLLEN, K. A. & BARB, K. H. 1981. Pearson's r and coarsely categorized measures. *American Sociological Review***,** 232-239.

BURTON, A., ALTMAN, D. G., ROYSTON, P. & HOLDER, R. L. 2006. The design of simulation studies in medical statistics. *Statistics in medicine,* 25**,** 4279-4292.

BURZYKOWSKI, T. & BUYSE, M. 2006. Surrogate threshold effect: an alternative measure for meta-analytic surrogate endpoint validation. *Pharmaceutical statistics,* 5**,** 173-186.

BURZYKOWSKI, T., MOLENBERGHS, G. & BUYSE, M. 2003. The validation of surrogate end points by using data from randomized clinical trials: a case-study in advanced colorectal cancer. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 167**,** 103-124.

BURZYKOWSKI, T., MOLENBERGHS, G. & BUYSE, M. 2004. The validation of surrogate end points by using data from randomized clinical trials: a case-

study in advanced colorectal cancer. *Journal of the Royal Statistical Society Series a-Statistics in Society,* 167**,** 103-124.

BURZYKOWSKI, T., MOLENBERGHS, G. & BUYSE, M. 2005. *The evaluation of surrogate endpoints*, Springer.

BURZYKOWSKI, T., MOLENBERGHS, G., BUYSE, M., GEYS, H. & RENARD, D. 2001. Validation of surrogate end points in multiple randomized clinical trials with failure time end points. *Journal of the Royal Statistical Society: Series C (Applied Statistics),* 50**,** 405-422.

BUYSE, M. & MOLENBERGHS, G. 1998. Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics***,** 1014-1029.

BUYSE, M., MOLENBERGHS, G. & BURZYKOWSKI, T. 2000a. Correction to 'Criteria for the validation of surrogate endpoints in randomized experiments' by M. Buyse and G. Molenberghs (Biometrics 1998;54:1014-1029). *Biometrics,* 56**,** 324-325.

BUYSE, M., MOLENBERGHS, G., BURZYKOWSKI, T., RENARD, D. & GEYS, H. 2000b. The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics,* 1**,** 49-67.

CHEN, C., WANG, H. W. & SNAPINN, S. M. 2003a. Proportion of treatment effect (PTE) explained by a surrogate marker. *Statistics in medicine,* 22**,** 3449-3459.

CHEN, H., GENG, Z. & JIA, J. Z. 2007. Criteria for surrogate end points. *Journal of the Royal Statistical Society Series B-Statistical Methodology,* 69**,** 919-932.

CHEN, S. X., LEUNG, D. H. Y. & QIN, J. 2003b. Information recovery in a study with surrogate endpoints. *Journal of the American Statistical Association,* 98**,** 1052-1062.

CHEN, S. X., LEUNG, D. H. Y. & QIN, J. 2008. Improving semiparametric estimation by using surrogate data. *Journal of the Royal Statistical Society Series B-Statistical Methodology,* 70**,** 803-823.

CLOTS 2013. Effectiveness of intermittent pneumatic compression in reduction of risk of deep vein thrombosis in patients who have had a stroke (CLOTS 3): a multicentre randomised controlled trial. *Lancet,* 382, 516-524.

COCHRAN, W. G. 1968. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics***,** 295-313.

COMMITTEE, H. O. C. H. 2004. The prevention of venous thromboembolism in hospitalised patients. *Second report of session,* 5**,** 2007.

CONLON, A., TAYLOR, J. M. & ELLIOTT, M. R. 2013. Surrogacy Assessment Using Principal Stratification When Surrogate and Outcome Measures are Multivariate Normal. *Biostatistics,* 15 (2), 266-283.

COWLES, M. K. 2002. Bayesian estimation of the proportion of treatment effect captured by a surrogate marker. *Statistics in medicine,* 21**,** 811-834.

COX, D. R. 1957. Note on grouping. *Journal of the American Statistical Association,* 52**,** 543-547.

COX, D. R. & WERMUTH, N. 2004. Causality: A statistical view. *International Statistical Review,* 72**,** 285-305.

DAI, J. Y. & HUGHES, J. P. 2012. A unified procedure for meta-analytic evaluation of surrogate end points in randomized clinical trials. *Biostatistics,* 13**,** 609-624.

DAY, N. & DUFFY, S. 1996. Trial Design Based on Surrogate End Points--Application to Comparison of Different Breast Screening Frequencies. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 49-60.

DESLANDES, E. & CHEVRET, S. 2007. Assessing surrogacy from the joint modelling of multivariate longitudinal data and survival: application to clinical trial data on chronic lymphocytic leukaemia. *Statistics in medicine,* 26**,** 5411-5421.

DIBAJ, S., FAGHIHZADEH, S. & JALAIE, S. 2010. Mistake on quantifying the indirect treatment effect via surrogate markers. Letter to the editor about paper by Y. Qu and M. Case (Statistics in Medicine 2006; 25:223-231). *Statistics in medicine,* 29**,** 2067.

DITLEVSEN, S., CHRISTENSEN, U., LYNCH, J., DAMSGAARD, M. T. & KEIDING, N. 2005a. The mediation proportion: a structural equation approach for estimating the proportion of exposure effect on outcome explained by an intermediate variable. *Epidemiology,* 16**,** 114-120.

DITLEVSEN, S., KEIDING, N., CHRISTENSEN, U., DAMSGAARD, M. T. & LYNCH, J. 2005b.Mediation proportion. Letter to the editor about paper by S. Ditlevsen, U. Christensen, J. Lynch, et al (Epidemiology 2005; 16:114-120). *Epidemiology,* 16**,** 592.

DUNNING, A. J. 2008. Comment on 'Evaluating a surrogate endpoint at three levels, with application to vaccine development'. Letter to the editor about paper by P. B. Gilbert, L. Qin and S. G. Self (Statistics in Medicine 2008; 27:4758-4778). *Statistics in medicine,* 27**,** 6268-6270.

ELLIOTT, M. R., CONLON, A. S., LI, Y., KACIROTI, N. & TAYLOR, J. M. 2014. Surrogacy marker paradox measures in meta-analytic settings. *Biostatistics***,** 16 (2), 400-412.

ELLIOTT, M. R., LI, Y. & TAYLOR, J. M. G. 2013. Accommodating missingness when assessing surrogacy via principal stratification. *Clinical Trials,* 10**,** 363-377.

EMSLEY, R., DUNN, G. & WHITE, I. R. 2010. Mediation and moderation of treatment effects in randomised controlled trials of complex interventions. *Statistical methods in medical research,* 19**,** 237-270.

ENSOR, H., LEE, R. J., SUDLOW, C. & WEIR, C. J. 2015. Statistical approaches for evaluating surrogate outcomes in clinical trials: a systematic review. *Journal of biopharmaceutical statistics,* Pre-published online.

FIRTH, D. 1993. Bias reduction of maximum likelihood estimates. *Biometrika,* 80**,** 27-38.

FLEMING, T. R., PRENTICE, R. L., PEPE, M. S. & GLIDDEN, D. 1994. Surrogate and auxiliary endpoints in clinical trials, with potential applications in cancer and AIDS research. *Statistics in medicine,* 13**,** 955-968.

FOLLMANN, D. 2006. Augmented designs to assess immune response in vaccine trials. *Biometrics,* 62**,** 1161-1169.

FRANGAKIS, C. E. & RUBIN, D. B. 2004. Principal stratification in causal inference. *Biometrics,* 58**,** 21-29.

FREEDMAN, L. 2005. Commentary on 'Assessing surrogates as trial endpoints using mixed models'. Commentary on paper by E. L. Korn, P. S. Albert and L. M. McShane (Statistics in Medicine 2005; 24:163-182). *Statistics in medicine,* 24**,** 183-185.

FREEDMAN, L. S., GRAUBARD, B. I. & SCHATZKIN, A. 1992. Statistical validation of intermediate endpoints for chronic diseases. *Statistics in medicine,* 11**,** 167-178.

GABRIEL, E. E. & GILBERT, P. B. 2014. Evaluating principal surrogate endpoints with time-to-event data accounting for time-varying treatment efficacy. *Biostatistics,* 15**,** 251-265.

GABRIEL, E. E., SACHS, M. C. & GILBERT, P. B. 2015. Comparing and combining biomarkers as principle surrogates for time-to-event clinical endpoints. *Statistics in medicine,* 34**,** 381-395.

GHOSH, D. 2008a. Composite endpoint analysis for assessing surrogacy with censored data. *Technical Report, Department of Statistics, Penn State University*.

GHOSH, D. 2008b. Semiparametric inference for surrogate endpoints with bivariate censored data. *Biometrics,* 64**,** 149-156.

GHOSH, D. 2009. On assessing surrogacy in a single trial setting using a semicompeting risks paradigm. *Biometrics,* 65**,** 521-529.

GHOSH, D., ELLIOTT, M. R. & TAYLOR, J. M. G. 2010. Links between analysis of surrogate endpoints and endogeneity. *Statistics in medicine,* 29**,** 2869-2879.

GHOSH, D., TAYLOR, J. M. G. & SARGENT, D. J. 2012a. Meta-analysis for Surrogacy: Accelerated Failure Time Models and Semicompeting Risks Modeling. *Biometrics,* 68**,** 226-232.

GHOSH, D., TAYLOR, J. M. G. & SARGENT, D. J. 2012b. Rejoinder for "Meta-analysis for Surrogacy: Accelerated Failure Time Models and Semicompeting Risks Modeling". *Biometrics,* 68**,** 245-247.

GILBERT, P. B., BOSCH, R. J. & HUDGENS, M. G. 2003. Sensitivity analysis for the assessment of causal vaccine effects on viral load in HIV vaccine trials. *Biometrics,* 59**,** 531-541.

GILBERT, P. B., GABRIEL, E. E., HUANG, Y. & CHAN, I. S. 2015. Surrogate Endpoint Evaluation: Principal Stratification Criteria and the Prentice Definition. *Journal of Causal Inference,* 3 (2), 157-175.

GILBERT, P. B. & HUDGENS, M. 2006. Evaluating causal effect predictiveness of candidate surrogate endpoints. *UW Biostatistics Working Paper Series,* Paper 291.

GILBERT, P. B. & HUDGENS, M. G. 2008. Evaluating candidate principal surrogate endpoints. *Biometrics,* 64**,** 1146-1154.

GILBERT, P. B., HUDGENS, M. G. & WOLFSON, J. 2011. Commentary on "Principal stratification - a goal or a tool?" by Judea Pearl. *The international journal of biostatistics,* 7 (1), 1-15.

GILBERT, P. B., QIN, L. & SELF, S. G. 2008. Evaluating a surrogate endpoint at three levels, with application to vaccine development. *Statistics in medicine,* 27**,** 4758-4778.

GILBERT, P. B., QIN, L. & SELF, S. G. 2009. Response to Andrew Dunning's comment on 'Evaluating a surrogate endpoint at three levels, with application to vaccine development'. Authors' reply to letter to the editor by A. J. Dunning (Statistics in Medicine 2008; 27:6268-6270). *Statistics in medicine,* 28**,** 716-719.

HEINZE, G. & SCHEMPER, M. 2002. A solution to the problem of separation in logistic regression. *Statistics in medicine,* 21**,** 2409-2419.

HERNANDEZ, J. C. Vicente-Villardon, J. L. 2013. OrdinalLogisticBiplot: Biplot representations of ordinal variables. *R.*

HUANG, J. & HUANG, B. 2010. Evaluating the Proportion of Treatment Effect Explained by a Continuous Surrogate Marker in Logistic or Probit Regression Models. *Statistics in Biopharmaceutical Research,* 2**,** 229-238.

HUANG, Y. & GILBERT, P. B. 2011. Comparing biomarkers as principal surrogate endpoints. *Biometrics,* 67**,** 1442-51.

HUANG, Y., GILBERT, P. B. & WOLFSON, J. 2013. Design and Estimation for Evaluating Principal Surrogate Markers in Vaccine Trials. *Biometrics,* 69**,** 301-309.

I-BIOSTAT. 2015. *Interuniversity Institute for Biostatistics and statistical Bioinformatics* [Online]. I-Biostat. Available: http://ibiostat.be/online-resources [Accessed 2015].

JOFFE, M. 2011. Principal Stratification and Attribution Prohibition: Good Ideas Taken Too Far. *International Journal of Biostatistics,* 7, 1-22.

JOFFE, M. M. & GREENE, T. 2009. Related causal frameworks for surrogate outcomes. *Biometrics,* 65**,** 530-538.

JOHNSON, K. R., FREEMANTLE, N., ANTHONY, D. M. & LASSERE, M. N. D. 2009. LDL-cholesterol differences predicted survival benefit in statin trials by the surrogate threshold effect (STE). *Journal of Clinical Epidemiology,* 62**,** 328-336.

JU, C. A. & GENG, Z. 2010. Criteria for surrogate end points based on causal distributions. *Journal of the Royal Statistical Society Series B-Statistical Methodology,* 72**,** 129-142.

KAPLAN, E. L. & MEIER, P. 1958. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association,* 53**,** 457-481.

KASSAI, B., SHAH, N. R., LEIZOROVICZA, A., CUCHERAT, M., GUEYFFIER, F. & BOISSEL, J. P. 2005. The true treatment benefit is unpredictable in clinical trials using surrogate outcome measured with diagnostic tests. *Journal of Clinical Epidemiology,* 58**,** 1042-1051.

KAUFMAN, J. S., MACLEHOSE, R. F., KAUFMAN, S. & GREENLAND, S. 2005. The mediation proportion. Letter to the editor about paper by S. Ditlevsen, N. Keiding, U. Christensen, et al (Epidemiology 2005; 16:114-120). *Epidemiology,* 16**,** 710.

KENT, J. T. 1983. Information gain and a general measure of correlation. *Biometrika,* 70**,** 163-173.

KORN, E. L., ALBERT, P. S. & MCSHANE, L. M. 2005a. Assessing surrogates as trial endpoints using mixed models. *Statistics in medicine,* 24**,** 163-182.

KORN, E. L., ALBERT, P. S. & MCSHANE, L. M. 2005b. Rejoinder to commentary by Dr Freedman of 'Assessing surrogates as trial endpoints using mixed models'. Rejoinder to commentary by L. Freedman (Statistics in Medicine 2005; 24:183-185). *Statistics in medicine,* 24**,** 187-190.

KRIEG, E. F. 1999. Biases induced by coarse measurement scales. *Educational and Psychological Measurement,* 59**,** 749-766.

KUROKI, M. 2013. Sharp bounds on causal effects using a surrogate endpoint. *Statistics in Medicine,* 32**,** 4338-4347.

LASSERE, M., JOHNSON, K., HUGHES, M., ALTMAN, D., BUYSE, M., GALBRAITH, S. & WELLS, G. 2007a. Simulation studies of surrogate endpoint validation using single trial and multitrial statistical approaches. *The Journal of rheumatology,* 34**,** 616-619.

LASSERE, M. N. 2008. The Biomarker-Surrogacy Evaluation Schema: a review of the biomarker-surrogate literature and a proposal for a criterion-based, quantitative, multidimensional hierarchical levels of evidence schema for evaluating the status of biomarkers as surrogate endpoints. *Statistical methods in medical research,* 17**,** 303-340.

LASSERE, M. N., JOHNSON, K. R., BOERS, M., TUGWELL, P., BROOKS, P., SIMON, L., STRAND, V., CONAGHAN, P. G., OSTERGAARD, M., MAKSYMOWYCH, W. P., LANDEWE, R., BRESNIHAN, B., TAK, P. P., WAKEFIELD, R., MEASE, P., BINGHAM, C. O., HUGHES, M., ALTMAN, D., BUYSE, M., GALBRAITH, S. & WELLS, G. 2007b. Definitions and validation criteria for biomarkers and surrogate endpoints: development and testing of a quantitative hierarchical levels of evidence schema. *Journal of Rheumatology,* 34**,** 607-615.

LAURITZEN, S. L. 2004. Discussion on causality. *Scandinavian Journal of Statistics,* 31**,** 189-193.

LE CESSIE, S. & VAN HOUWELINGEN, J. C. 1992. Ridge estimators in logistic regression. *Applied statistics***,** 41 (1), 191-201.

LI, W. & QU, Y. M. 2010. Adjustment for the measurement error in evaluating biomarkers. *Statistics in medicine,* 29**,** 2338-2346.

LI, Y., TAYLOR, J. M. G. & ELLIOTT, M. R. 2010. A Bayesian approach to surrogacy assessment using principal stratification in clinical trials. *Biometrics,* 66**,** 523-531.

LI, Y., TAYLOR, J. M. G., ELLIOTT, M. R. & SARGENT, D. J. 2011. Causal assessment of surrogacy in a meta-analysis of colorectal cancer trials. *Biostatistics (Oxford, England),* 12**,** 478-92.

LIU, W., ZHANG, B., ZHANG, H. & ZHANG, Z. 2014. Likelihood-based methods for evaluating principal surrogacy in augmented vaccine trials. *Statistical methods in medical research***,** Doi: 10/1177/0962280214565833.

MACKINNON, D. P., LOCKWOOD, C. M., BROWN, C. H., WANG, W. & HOFFMAN, J. M. 2007. The intermediate endpoint effect in logistic and probit regression. *Clinical Trials,* 4**,** 499-513.

MCCULLOCH, C. E. & SEARLE, S. R. 2001. Linear Mixed Models (LMMs). *Generalized, linear, and mixed models***,** 156-186. Online: Wiley.

MCHUGH, G. S., BUTCHER, I., STEYERBERG, E. W., MARMAROU, A., LU, J., LINGSMA, H. F., WEIR, J., MAAS, A. I. & MURRAY, G. D. 2010b. A simulation study evaluating approaches to the analysis of ordinal outcome data in randomized controlled trials in traumatic brain injury: results from the IMPACT Project. *Clinical Trials,* 7**,** 44-57.

MEALLI, F. & MATTEI, A. 2012. A refreshing account of principal stratification. *The international journal of biostatistics,* 8.

MIAO, X., LI, X., GILBERT, P. B. & CHAN, I. S. 2013. A Multiple Imputation Approach for the Evaluation of Surrogate Markers in the Principal Stratification Causal Inference Framework. *Risk Assessment and Evaluation of Predictions*. Springer.

MIAO, X. P., WANG, Y. C. & GANGOPADHYAY, A. 2012. An entropy-based nonparametric test for the validation of surrogate endpoints. *Statistics in medicine,* 31**,** 1517-1530.

MOLENBERGHS, G. 2012. Discussion Contribution to 091037PR4 (Ghosh, Taylor, and Sargent). *Biometrics,* 68**,** 233-235.

MOLENBERGHS, G., BURZYKOWSKI, T., ALONSO, A. & BUYSE, M. 2004. A perspective on surrogate endpoints in controlled clinical trials. *Statistical methods in medical research,* 13**,** 177-206.

MOLENBERGHS, G., GEYS, H. & BUYSE, M. 2001. Evaluation of surrogate endpoints in randomized experiments with mixed discrete and continuous outcomes. *Statistics in medicine,* 20**,** 3023-3038.

MOLENBERGHS, G., VERBEKE, G. & IDDI, S. 2011. Pseudo-likelihood methodology for partitioned large and complex samples. *Statistics & Probability Letters,* 81**,** 892-901.

MURRAY, C. J. & LOPEZ, A. D. 1996. *Global burden of disease,* Harvard University Press Cambridge, MA.

O'QUIGLEY, J. & FLANDRE, P. 2006. Quantification of the Prentice criteria for surrogate endpoints. *Biometrics,* 62**,** 297-300.

PEARL, J. 2011. Principal stratification--a goal or a tool? *The international journal of biostatistics,* 7**,** 20.

PEARL, J. & BAREINBOIM, E. 2011. Transportability across studies: A formal approach. DTIC Document.

PLACKETT, R. L. 1965. A class of bivariate distributions. *Journal of the American Statistical Association,* 60**,** 516-522.

PRENTICE, R. L. 1989. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in medicine,* 8**,** 431-440.

PRYSELEY, A., TILAHUN, A., ALONSO, A. & MOLENBERGHS, G. 2007. Information-theory based surrogate marker evaluation from several randomized clinical trials with continuous true and binary surrogate endpoints. *Clinical Trials,* 4**,** 587-597.

PRYSELEY, A., TILAHUN, A., ALONSO, A. & MOLENBERGHS, G. 2010. Using earlier measures in a longitudinal sequence as a potential surrogate for a later one. *Computational statistics & data analysis,* 54**,** 1342-1354.

PRYSELEY, A., TILAHUN, A., ALONSO, A. & MOLENBERGHS, G. 2011. An information-theoretic approach to surrogate-marker evaluation with failure time endpoints. *Lifetime data analysis,* 17**,** 195-214.

QIN, L., GILBERT, P. B., COREY, L., MCELRATH, M. J. & SELF, S. G. 2007. A framework for assessing immunological correlates of protection in vaccine trials. *Journal of Infectious Diseases,* 196**,** 1304-1312.

QIN, L., GILBERT, P. B., FOLLMANN, D. & LI, D. F. 2008. Assessing surrogate endpoints in vaccine trials with case-cohort sampling and the Cox model. *Annals of Applied Statistics,* 2**,** 386-407.

QU, Y. & CASE, M. 2007. Quantifying the effect of the surrogate marker by information gain. *Biometrics,* 63**,** 958-962.

QU, Y. M. & CASE, M. 2006. Quantifying the indirect treatment effect via surrogate markers. *Statistics in medicine,* 25**,** 223-231.

QU, Y. M. & CASE, M. 2010. Reply to 'Mistake on quantifying the indirect treatment effect via surrogate markers'. Authors' reply to letter to the editor

by S. Dibaj, S. Faghihzadeh and S. Jalaie (Statistics in Medicine 2010; 29:2067). *Statistics in medicine,* 29**,** 2068.

R CORE TEAM (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

RENARD, D., GEYS, H., MOLENBERGHS, G., BURZYKOWSKI, T. & BUYSE, M. 2002. Validation of surrogate endpoints in multiple randomized clinical trials with discrete outcomes. *Biometrical Journal,* 44**,** 921-935.

RENARD, D., GEYS, H., MOLENBERGHS, G., BURZYKOWSKI, T., BUYSE, M. & VANGENEUGDEN, T. 2003. Validation of a longitudinally measured surrogate marker for a time-to-event endpoint. *Journal of Applied Statistics,* 30**,** 235-247.

RENFRO, L. A., SHI, Q., SARGENT, D. J. & CARLIN, B. P. 2012. Bayesian adjusted R2 for the meta-analytic evaluation of surrogate time-to-event endpoints in clinical trials. *Statistics in medicine,* 31**,** 743-61.

RENFRO, L. A., SHI, Q., XUE, Y., LI, J., SHANG, H. & SARGENT, D. J. 2014. Center-within-trial versus trial-level evaluation of surrogate endpoints. *Computational statistics & data analysis,* 78**,** 1-20.

ROBINS, J. M. & GREENLAND, S. 1992. Identifiability and exchangeability for direct and indirect effects. *Epidemiology,* 3**,** 143-155.

RUBIN, D. B. 2004. Direct and Indirect Causal Effects via Potential Outcomes*. *Scandinavian Journal of Statistics,* 31**,** 161-170.

SANDERCOCK, P., LINDLEY, R., WARDLAW, J., DENNIS, M., LEWIS, S., VENABLES, G., KOBAYASHI, A., CZLONKOWSKA, A., BERGE, E. & SLOT, K. B. 2008. The third international stroke trial (IST-3) of thrombolysis for acute ischaemic stroke. *Trials,* 9**,** 37.

SARKAR, S. & QU, Y. M. 2007. Quantifying the treatment effect explained by markers in the presence of measurement error. *Statistics in medicine,* 26**,** 1955-1963.

SHANNON, C. E. & WEAVER, W. 1948. A mathematical theory of communication. American Telephone and Telegraph Company.

SHI, Q., RENFRO, L. A., BOT, B. M., BURZYKOWSKI, T., BUYSE, M. & SARGENT, D. J. 2011. Comparative assessment of trial-level surrogacy measures for candidate time-to-event surrogate endpoints in clinical trials. *Computational statistics & data analysis,* 55**,** 2748-2757.

SLOT, K. B., BERGE, E., DORMAN, P., LEWIS, S., DENNIS, M. & SANDERCOCK, P. 2008. Impact of functional status at six months on long term survival in patients with ischaemic stroke: prospective cohort studies. *BMJ,* 336**,** 376-379.

TAYLOR, A. B., WEST, S. G. & AIKEN, L. S. 2006. Loss of power in logistic, ordinal logistic, and probit regression when an outcome variable is coarsely categorized. *Educational and Psychological Measurement,* 66**,** 228-239.

TAYLOR, J. M. & YU, M. 2002. Bias and efficiency loss due to categorizing an explanatory variable. *Journal of Multivariate Analysis,* 83**,** 248-263.

TAYLOR, J. M. G., WANG, Y. & THIEBAUT, R. 2005. Counterfactual links to the proportion of treatment effect explained by a surrogate marker. *Biometrics,* 61**,** 1102-1111.

TEMPLE, R. 1999 "Are surrogate markers adequate to assess cardiovascular disease drugs?." *Journal of the American Medical Association* 282.8, 790-795.

THOMPSON, L. 2009. *R (and S-PLUS) Manual to Accompany, Agresti's Categorical Data Analysis (2002), 2nd edition* [Online]. Available: https://home.comcast.net/~lthompson221/Splusdiscrete2.pdf [Accessed 16/12/2014 2014].

TIBALDI, F., BARBOSA, F. T. & MOLENBERGHS, G. 2004. Modelling associations between time-to-event responses in pilot cancer clinical trials using a Plackett-Dale model. *Statistics in medicine,* 23**,** 2173-2186.

TIBALDI, F. S., ABRAHANTES, J. C., MOLENBERGHS, G., RENARD, D., BURZYKOWSKI, T., BUYSE, M., PARMAR, M., STIJNEN, T. & WOLFINGER, R. 2003b. Simplified hierarchical linear models for the evaluation of surrogate endpoints. *Journal of Statistical Computation and Simulation,* 73**,** 643-658.

TILAHUN, A., MARINGWA, J., GEYS, H., ALONSO, A., RAEYMAEKERS, L., MOLENBERGHS, G., KIEBOOM, G. V. D., DRINKENBURG, P. & BIJNENS, L. 2009. Investigating association between behavior, corticosterone, heart rate, and blood pressure in rats using surrogate marker evaluation methodology. *Journal of biopharmaceutical statistics,* 19**,** 133-149.

TILAHUN, A., PRYSELEY, A., ALONSO, A. & MOLENBERGHS, G. 2007. Flexible surrogate marker evaluation from several randomized clinical trials with continuous endpoints, using R and SAS. *Computational statistics & data analysis,* 51**,** 4152-4163.

TILAHUN, A., PRYSELEY, A., ALONSO, A. & MOLENBERGHS, G. 2008b. Information Theory–Based Surrogate Marker Evaluation from Several Randomized Clinical Trials with Binary Endpoints, Using SAS. *Journal of biopharmaceutical statistics,* 18**,** 326-341.

VAN SWIETEN, J., KOUDSTAAL, P., VISSER, M., SCHOUTEN, H. & VAN GIJN, J. 1988. Interobserver agreement for the assessment of handicap in stroke patients. *Stroke,* 19**,** 604-607.

VAN WALRAVEN, C., OAKE, N., COYLE, D., TALJAARD, M. & FORSTER, A. J. 2009. Changes in surrogate outcomes can be translated into clinical outcomes using a Monte Carlo model. *Journal of Clinical Epidemiology,* 62**,** 1306-1315.

VANDERWEELE, T. J. 2013. Surrogate measures and consistent surrogates. *Biometrics,* 69**,** 561-569.

VICENTE VILLARDÓN, J. L. 2015. *RE: Personal communication.*

WANG, Y., MOGG, R. & LUNCEFORD, J. 2012. Evaluating correlation-based metric for surrogate marker qualification within a causal correlation framework. *Biometrics,* 68**,** 617-27.

WANG, Y. & TAYLOR, J. M. 2002. A measure of the proportion of treatment effect explained by a surrogate marker. *Biometrics,* 58**,** 803-812.

WEIR, C. J. & WALLEY, R. J. 2006. Statistical evaluation of biomarkers as surrogate endpoints: a literature review. *Statistics in medicine,* 25**,** 183-203.

WINSHIP, C. & MARE, R. D. 1984. Regression models with ordinal variables. *American Sociological Review***,** 512-525.

WOLFSON, J. & GILBERT, P. 2010. Statistical identifiability and the surrogate endpoint problem, with application to vaccine trials. *Biometrics,* 66**,** 1153-1161.

WOLFSON, J. & HENN, L. 2014. Hard, harder, hardest: principal stratification, statistical identifiability, and the inherent difficulty of finding surrogate endpoints. *Emerging themes in epidemiology,* 11**:**14.

WU, Z., HE, P. & GENG, Z. 2011. Sufficient conditions for concluding surrogacy based on observed data. *Statistics in medicine,* 30**,** 2422-2434.

ZIGLER, C. M. & BELIN, T. R. 2012. A Bayesian Approach to Improved Estimation of Causal Effect Predictiveness for a Principal Surrogate Endpoint. *Biometrics,* 68**,** 922-932.

Appendix A: Binary-ordinal

Tables presented here are for a binary surrogate and ordinal true outcome, for each scenario covered in the simulation for the binary-ordinal setting. The results given are the median R^2 values at the trial and individual level for each scenario, the IQR, the coverage of the confidence intervals and the median 95% confidence interval limits.

No. trials	Trial Size	R_h^2	IQR $\mathbf{R_h^2}$	% Cover CIs	Low CI	Upp CI
5	10	0.347	0.154	100%	0.046	0.757
5	20	0.318	0.127	100%	0.062	0.650
5	40	0.307	0.093	100%	0.103	0.562
5	60	0.305	0.086	100%	0.126	0.519
5	100	0.300	0.063	99%	0.155	0.465
5	150	0.311	0.058	99%	0.185	0.446
5	200	0.307	0.062	98%	0.198	0.426
5	300	0.304	0.058	96%	0.213	0.401
10	10	0.337	0.107	100%	0.046	0.742
10	20	0.310	0.071	100%	0.066	0.641
10	40	0.298	0.069	100%	0.096	0.547
10	60	0.293	0.063	100%	0.121	0.499
10	100	0.302	0.056	100%	0.156	0.467
10	150	0.298	0.051	100%	0.178	0.433
10	200	0.300	0.045	99%	0.194	0.417
10	300	0.294	0.040	99%	0.207	0.389
20	10	0.342	0.072	100%	0.046	0.743
20	20	0.301	0.060	100%	0.064	0.629
20	40	0.302	0.044	100%	0.098	0.547
20	60	0.297	0.039	100%	0.121	0.501
20	100	0.293	0.036	100%	0.151	0.455
20	150	0.294	0.034	100%	0.173	0.427
20	200	0.295	0.031	100%	0.188	0.411
20	300	0.292	0.031	100%	0.205	0.386
30	10	0.340	0.071	100%	0.047	0.739
30	20	0.307	0.044	100%	0.064	0.628
30	40	0.301	0.040	100%	0.100	0.545
30	60	0.295	0.033	100%	0.121	0.500
30	100	0.294	0.027	100%	0.151	0.454
30	150	0.297	0.024	100%	0.176	0.429
30	200	0.293	0.023	100%	0.187	0.408
30	300	0.293	0.025	100%	0.206	0.387

Table A.1*: Simulation binary-ordinal: R_h^2 results. Where R_h^2=0.64 and R_{ht}^2=0.90 and the odds are proportional*

No. trials	Trial Size	R^2_{ht}	IQR R^2_{ht}	% Cover CI	Low CI	Upp CI
5	10	0.781	0.367	82%	0.120	0.988
5	20	0.850	0.207	90%	0.222	0.994
5	40	0.908	0.184	88%	0.359	0.997
5	60	0.923	0.146	93%	0.408	0.998
5	100	0.924	0.135	92%	0.411	0.998
5	150	0.947	0.114	91%	0.502	0.999
5	200	0.939	0.115	93%	0.468	0.998
5	300	0.949	0.111	93%	0.513	0.999
10	10	0.616	0.304	68%	0.121	0.922
10	20	0.737	0.239	72%	0.249	0.957
10	40	0.805	0.175	76%	0.352	0.973
10	60	0.838	0.153	84%	0.412	0.979
10	100	0.862	0.132	91%	0.462	0.984
10	150	0.869	0.127	94%	0.479	0.985
10	200	0.885	0.119	93%	0.516	0.987
10	300	0.900	0.090	95%	0.553	0.990
20	10	0.571	0.203	45%	0.207	0.843
20	20	0.681	0.175	47%	0.328	0.897
20	40	0.792	0.125	70%	0.485	0.943
20	60	0.803	0.120	76%	0.503	0.947
20	100	0.831	0.098	86%	0.552	0.957
20	150	0.840	0.104	86%	0.568	0.960
20	200	0.861	0.083	92%	0.606	0.966
20	300	0.870	0.082	95%	0.625	0.969
30	10	0.549	0.169	21%	0.249	0.791
30	20	0.671	0.126	30%	0.384	0.864
30	40	0.766	0.116	53%	0.512	0.913
30	60	0.798	0.100	69%	0.560	0.928
30	100	0.829	0.092	83%	0.611	0.942
30	150	0.846	0.078	85%	0.640	0.949
30	200	0.849	0.069	85%	0.645	0.950
30	300	0.865	0.065	94%	0.672	0.957

Table A.2: *Simulation binary-ordinal: R^2_{ht}results, Where R^2_h=0.64 and R^2_{ht}=0.90 and the odds are proportional*

No. trials	Trial Size	R_h^2	IQR R_h^2	% Cover CIs	Low CI	Upp CI
5	10	0.336	0.172	100%	0.042	0.745
5	20	0.315	0.130	100%	0.063	0.643
5	40	0.301	0.096	100%	0.100	0.553
5	60	0.302	0.078	100%	0.120	0.516
5	100	0.304	0.064	99%	0.158	0.471
5	150	0.302	0.060	99%	0.177	0.440
5	200	0.297	0.065	98%	0.188	0.415
5	300	0.301	0.056	97%	0.211	0.396
10	10	0.330	0.106	100%	0.043	0.742
10	20	0.301	0.079	100%	0.060	0.635
10	40	0.299	0.060	100%	0.096	0.547
10	60	0.300	0.058	100%	0.121	0.507
10	100	0.294	0.052	100%	0.152	0.457
10	150	0.294	0.047	100%	0.173	0.428
10	200	0.294	0.039	100%	0.187	0.411
10	300	0.298	0.046	99%	0.209	0.394
20	10	0.327	0.082	100%	0.046	0.736
20	20	0.303	0.059	100%	0.065	0.629
20	40	0.298	0.041	100%	0.097	0.545
20	60	0.292	0.033	100%	0.119	0.497
20	100	0.291	0.039	100%	0.148	0.452
20	150	0.295	0.033	100%	0.174	0.428
20	200	0.292	0.027	100%	0.187	0.406
20	300	0.291	0.033	100%	0.203	0.384
30	10	0.327	0.073	100%	0.045	0.730
30	20	0.304	0.046	100%	0.062	0.628
30	40	0.293	0.038	100%	0.094	0.539
30	60	0.290	0.031	100%	0.118	0.495
30	100	0.290	0.031	100%	0.147	0.451
30	150	0.291	0.027	100%	0.172	0.425
30	200	0.290	0.025	100%	0.185	0.405
30	300	0.288	0.025	100%	0.202	0.381

Table A.3: *Simulation binary-ordinal: R_h^2 results. Where R_h^2=0.64 and R_{ht}^2=0.30 and the odds are proportional*

No. trials	Trial Size	R^2_{ht}	IQR R^2_{ht}	% Cover CIs	Low CI	Upp CI
5	10	0.672	0.485	70%	0.039	0.976
5	20	0.670	0.446	74%	0.037	0.976
5	40	0.673	0.476	72%	0.039	0.976
5	60	0.637	0.490	72%	0.026	0.971
5	100	0.680	0.437	80%	0.042	0.977
5	150	0.617	0.482	80%	0.021	0.968
5	200	0.644	0.471	77%	0.028	0.972
5	300	0.612	0.471	80%	0.020	0.967
10	10	0.363	0.390	79%	0.008	0.811
10	20	0.386	0.308	87%	0.011	0.824
10	40	0.378	0.335	94%	0.010	0.820
10	60	0.414	0.358	93%	0.016	0.839
10	100	0.428	0.344	93%	0.020	0.846
10	150	0.404	0.329	93%	0.014	0.834
10	200	0.446	0.300	94%	0.025	0.854
10	300	0.374	0.353	97%	0.009	0.818
20	10	0.296	0.229	92%	0.024	0.656
20	20	0.326	0.194	99%	0.036	0.679
20	40	0.319	0.219	95%	0.033	0.674
20	60	0.339	0.246	98%	0.042	0.690
20	100	0.331	0.208	96%	0.038	0.683
20	150	0.329	0.219	95%	0.037	0.681
20	200	0.347	0.267	97%	0.045	0.696
20	300	0.343	0.247	96%	0.043	0.693
30	10	0.261	0.187	98%	0.036	0.564
30	20	0.306	0.181	95%	0.059	0.606
30	40	0.324	0.188	93%	0.069	0.620
30	60	0.315	0.182	91%	0.064	0.612
30	100	0.331	0.213	92%	0.073	0.628
30	150	0.349	0.220	92%	0.085	0.642
30	200	0.277	0.195	94%	0.044	0.577
30	300	0.340	0.202	95%	0.079	0.634

Table A.4: *Simulation binary-ordinal: R^2_{ht} results. Where R^2_h=0.64 and R^2_{ht}=0.30 and the odds are proportional*

No. trials	Trial Size	R_h^2	IQR R_h^2	% Cover CIs	Low CI	Upp CI
5	10	0.213	0.125	100%	0.012	0.655
5	20	0.174	0.104	100%	0.015	0.499
5	40	0.149	0.055	100%	0.022	0.387
5	60	0.147	0.056	100%	0.029	0.339
5	100	0.146	0.044	100%	0.044	0.293
5	150	0.137	0.043	100%	0.052	0.254
5	200	0.138	0.033	100%	0.061	0.238
5	300	0.140	0.028	99%	0.074	0.223
10	10	0.214	0.093	100%	0.016	0.651
10	20	0.161	0.060	100%	0.014	0.486
10	40	0.146	0.040	100%	0.022	0.376
10	60	0.140	0.040	100%	0.028	0.328
10	100	0.141	0.028	100%	0.042	0.286
10	150	0.137	0.027	100%	0.052	0.253
10	200	0.134	0.025	100%	0.059	0.233
10	300	0.134	0.022	100%	0.071	0.215
20	10	0.215	0.068	100%	0.019	0.647
20	20	0.165	0.040	100%	0.016	0.489
20	40	0.144	0.029	100%	0.023	0.372
20	60	0.142	0.030	100%	0.031	0.329
20	100	0.138	0.023	100%	0.042	0.282
20	150	0.136	0.017	100%	0.052	0.252
20	200	0.133	0.015	100%	0.059	0.232
20	300	0.133	0.017	100%	0.070	0.214
30	10	0.208	0.050	100%	0.018	0.646
30	20	0.165	0.032	100%	0.019	0.488
30	40	0.147	0.028	100%	0.023	0.378
30	60	0.143	0.022	100%	0.031	0.330
30	100	0.136	0.018	100%	0.041	0.279
30	150	0.134	0.018	100%	0.051	0.250
30	200	0.132	0.013	100%	0.058	0.230
30	300	0.133	0.013	100%	0.070	0.212

Table A.5: *Simulation binary-ordinal: R_h^2 results. Where R_h^2=0.30 and R_{ht}^2=0.90 and the odds are proportional*

No. trials	Trial Size	R^2_{ht}	IQR. R^2_{ht}	% Cover CIs	Low CI	Upp CI
5	10	0.753	0.325	85%	0.091	0.986
5	20	0.821	0.232	90%	0.173	0.992
5	40	0.866	0.222	86%	0.254	0.995
5	60	0.902	0.174	88%	0.341	0.997
5	100	0.913	0.154	93%	0.376	0.997
5	150	0.935	0.131	92%	0.453	0.998
5	200	0.937	0.099	94%	0.462	0.998
5	300	0.947	0.101	92%	0.504	0.999
10	10	0.531	0.320	74%	0.061	0.891
10	20	0.681	0.244	71%	0.183	0.943
10	40	0.785	0.190	81%	0.319	0.969
10	60	0.826	0.159	88%	0.389	0.977
10	100	0.852	0.148	85%	0.442	0.982
10	150	0.881	0.129	90%	0.506	0.987
10	200	0.885	0.115	93%	0.515	0.987
10	300	0.896	0.085	96%	0.543	0.989
20	10	0.451	0.226	38%	0.106	0.770
20	20	0.632	0.192	29%	0.271	0.874
20	40	0.726	0.139	56%	0.386	0.916
20	60	0.774	0.136	71%	0.457	0.936
20	100	0.817	0.104	83%	0.527	0.952
20	150	0.834	0.105	86%	0.557	0.958
20	200	0.849	0.106	92%	0.584	0.963
20	300	0.861	0.081	94%	0.607	0.966
30	10	0.447	0.194	17%	0.156	0.720
30	20	0.601	0.162	16%	0.303	0.823
30	40	0.714	0.126	34%	0.439	0.887
30	60	0.757	0.103	55%	0.499	0.909
30	100	0.798	0.088	79%	0.561	0.928
30	150	0.829	0.080	88%	0.611	0.942
30	200	0.838	0.089	85%	0.626	0.946
30	300	0.857	0.066	92%	0.658	0.953

Table A.6: *Simulation binary-ordinal: R^2_{ht} results. Where R^2_h=0.30 and R^2_{ht}=0.90 and the odds are proportional*

No. trials	Trial Size	R_h^2	IQR R_h^2	% Cover CIs	Low CI	Upp CI
5	10	0.216	0.142	100%	0.013	0.659
5	20	0.166	0.089	100%	0.014	0.493
5	40	0.154	0.064	100%	0.021	0.385
5	60	0.147	0.056	100%	0.030	0.336
5	100	0.135	0.043	100%	0.038	0.282
5	150	0.138	0.040	100%	0.052	0.256
5	200	0.137	0.039	99%	0.061	0.238
5	300	0.135	0.033	99%	0.070	0.216
10	10	0.211	0.101	100%	0.018	0.648
10	20	0.166	0.057	100%	0.015	0.493
10	40	0.151	0.048	100%	0.023	0.384
10	60	0.141	0.037	100%	0.029	0.327
10	100	0.136	0.032	100%	0.041	0.281
10	150	0.135	0.027	100%	0.051	0.250
10	200	0.131	0.025	100%	0.057	0.229
10	300	0.134	0.023	100%	0.070	0.214
20	10	0.212	0.070	100%	0.018	0.645
20	20	0.164	0.045	100%	0.017	0.489
20	40	0.145	0.032	100%	0.023	0.377
20	60	0.139	0.029	100%	0.030	0.325
20	100	0.136	0.021	100%	0.041	0.279
20	150	0.133	0.016	100%	0.051	0.248
20	200	0.134	0.018	100%	0.059	0.232
20	300	0.131	0.017	100%	0.068	0.210
30	10	0.209	0.050	100%	0.018	0.645
30	20	0.163	0.041	100%	0.016	0.488
30	40	0.145	0.027	100%	0.023	0.375
30	60	0.140	0.022	100%	0.030	0.328
30	100	0.136	0.019	100%	0.040	0.279
30	150	0.134	0.016	100%	0.051	0.249
30	200	0.132	0.015	100%	0.058	0.229
30	300	0.131	0.011	100%	0.068	0.210

Table A.7: *Simulation binary-ordinal: R_h^2 results: Where R_h^2=0.30 and R_{ht}^2=0.30 and the odds are proportional*

No. trials	Trial Size	R^2_{ht}	IQR R^2_{ht}	% Cover CIs	Low CI	Upp CI
5	10	0.561	0.539	74%	0.011	0.959
5	20	0.647	0.438	72%	0.029	0.973
5	40	0.653	0.402	79%	0.031	0.973
5	60	0.596	0.485	76%	0.016	0.965
5	100	0.670	0.491	71%	0.038	0.976
5	150	0.687	0.415	76%	0.046	0.978
5	200	0.638	0.426	80%	0.026	0.971
5	300	0.658	0.473	79%	0.033	0.974
10	10	0.327	0.326	76%	0.005	0.790
10	20	0.389	0.300	92%	0.012	0.826
10	40	0.400	0.381	96%	0.013	0.831
10	60	0.440	0.366	91%	0.023	0.852
10	100	0.431	0.360	95%	0.020	0.847
10	150	0.448	0.373	95%	0.025	0.855
10	200	0.415	0.385	94%	0.016	0.839
10	300	0.401	0.385	95%	0.014	0.832
20	10	0.238	0.227	92%	0.009	0.602
20	20	0.276	0.230	98%	0.017	0.637
20	40	0.327	0.239	98%	0.036	0.680
20	60	0.324	0.299	95%	0.035	0.678
20	100	0.329	0.202	97%	0.037	0.682
20	150	0.295	0.261	97%	0.024	0.654
20	200	0.332	0.228	96%	0.038	0.684
20	300	0.333	0.224	96%	0.039	0.685
30	10	0.211	0.171	97%	0.017	0.510
30	20	0.249	0.179	93%	0.031	0.550
30	40	0.281	0.192	82%	0.046	0.581
30	60	0.283	0.176	88%	0.047	0.583
30	100	0.303	0.154	92%	0.057	0.601
30	150	0.312	0.200	91%	0.062	0.610
30	200	0.302	0.208	95%	0.057	0.600
30	300	0.313	0.164	94%	0.063	0.610

Table A.8 : *Simulation binary-ordinal: R^2_{ht} results: Where R^2_h=0.30 and R^2_{ht}=0.30 and the odds are proportional*

No. trials	Trial Size	R_h^2	IQR R_h^2	% Cover CIs	Low CI	Upp CI
5	10	0.344	0.160	100%	0.040	0.754
5	20	0.318	0.115	100%	0.067	0.651
5	40	0.304	0.083	100%	0.096	0.555
5	60	0.310	0.087	99%	0.125	0.522
5	100	0.308	0.063	100%	0.160	0.474
5	150	0.304	0.064	98%	0.180	0.440
5	200	0.300	0.051	99%	0.193	0.419
5	300	0.302	0.058	95%	0.212	0.398
10	10	0.322	0.114	100%	0.044	0.731
10	20	0.311	0.084	100%	0.067	0.639
10	40	0.299	0.059	100%	0.097	0.548
10	60	0.295	0.051	100%	0.120	0.500
10	100	0.296	0.051	100%	0.152	0.459
10	150	0.296	0.049	100%	0.174	0.430
10	200	0.297	0.040	100%	0.189	0.414
10	300	0.288	0.050	100%	0.202	0.384
20	10	0.333	0.078	100%	0.047	0.738
20	20	0.304	0.056	100%	0.063	0.629
20	40	0.294	0.052	100%	0.096	0.542
20	60	0.294	0.038	100%	0.119	0.499
20	100	0.294	0.033	100%	0.150	0.454
20	150	0.293	0.032	100%	0.172	0.426
20	200	0.292	0.034	100%	0.186	0.408
20	300	0.292	0.031	100%	0.203	0.386
30	10	0.330	0.066	100%	0.047	0.739
30	20	0.297	0.051	100%	0.062	0.625
30	40	0.293	0.036	100%	0.095	0.539
30	60	0.292	0.036	100%	0.119	0.496
30	100	0.290	0.030	100%	0.150	0.451
30	150	0.292	0.030	100%	0.172	0.425
30	200	0.292	0.026	100%	0.187	0.408
30	300	0.288	0.021	100%	0.201	0.381

Table A.9: *Simulation binary-ordinal: R_h^2 results: Where R_h^2=0.64 and R_{ht}^2=0.90 and the odds are non-proportional*

No. trials	Trial Size	R^2_{ht}	IQR R^2_{ht}	% Cover CIs	Low CI	Upp CI
5	10	0.791	0.311	86%	0.132	0.989
5	20	0.856	0.275	90%	0.233	0.994
5	40	0.906	0.154	92%	0.354	0.997
5	60	0.916	0.152	88%	0.386	0.998
5	100	0.920	0.124	91%	0.399	0.998
5	150	0.936	0.124	93%	0.457	0.998
5	200	0.927	0.147	92%	0.423	0.998
5	300	0.945	0.114	93%	0.494	0.999
10	10	0.595	0.304	72%	0.103	0.915
10	20	0.748	0.226	69%	0.264	0.960
10	40	0.803	0.137	81%	0.349	0.972
10	60	0.837	0.133	86%	0.410	0.979
10	100	0.858	0.129	89%	0.453	0.983
10	150	0.881	0.113	92%	0.504	0.987
10	200	0.886	0.114	90%	0.517	0.987
10	300	0.881	0.123	93%	0.506	0.987
20	10	0.533	0.221	41%	0.171	0.821
20	20	0.683	0.161	41%	0.330	0.898
20	40	0.771	0.134	60%	0.452	0.935
20	60	0.808	0.120	79%	0.511	0.949
20	100	0.827	0.101	79%	0.544	0.955
20	150	0.844	0.093	89%	0.574	0.961
20	200	0.852	0.085	93%	0.590	0.964
20	300	0.858	0.080	94%	0.602	0.966
30	10	0.525	0.164	21%	0.224	0.777
30	20	0.668	0.138	30%	0.381	0.862
30	40	0.757	0.125	53%	0.499	0.909
30	60	0.788	0.098	64%	0.545	0.924
30	100	0.833	0.077	81%	0.617	0.943
30	150	0.829	0.078	87%	0.611	0.942
30	200	0.833	0.081	87%	0.618	0.944
30	300	0.858	0.080	94%	0.661	0.954

Table A.10: *Simulation binary-ordinal: R^2_{ht} results: Where R^2_h=0.64 and R^2_{ht}=0.90 and the odds are non-proportional*

No. trials	Trial Size	R_h^2	IQR R_h^2	% Cover CIs	Low CI	Upp CI
5	10	0.521	0.174	100%	0.109	0.856
5	20	0.521	0.163	98%	0.196	0.796
5	40	0.530	0.146	95%	0.291	0.746
5	60	0.531	0.155	93%	0.332	0.714
5	100	0.516	0.156	86%	0.368	0.648
5	150	0.530	0.136	81%	0.405	0.656
5	200	0.536	0.151	71%	0.426	0.642
5	300	0.490	0.155	53%	0.413	0.566
10	10	0.500	0.113	100%	0.108	0.840
10	20	0.505	0.109	100%	0.193	0.777
10	40	0.501	0.097	98%	0.277	0.706
10	60	0.502	0.098	99%	0.315	0.672
10	100	0.504	0.097	93%	0.358	0.634
10	150	0.499	0.099	88%	0.384	0.608
10	200	0.494	0.084	87%	0.394	0.590
10	300	0.488	0.106	70%	0.410	0.566
20	10	0.493	0.086	100%	0.107	0.836
20	20	0.501	0.074	100%	0.195	0.771
20	40	0.494	0.066	100%	0.274	0.694
20	60	0.501	0.068	100%	0.312	0.669
20	100	0.491	0.066	98%	0.351	0.624
20	150	0.495	0.071	98%	0.380	0.604
20	200	0.495	0.068	94%	0.396	0.590
20	300	0.489	0.066	88%	0.409	0.568
30	10	0.484	0.071	100%	0.105	0.823
30	20	0.496	0.061	100%	0.195	0.765
30	40	0.491	0.060	100%	0.270	0.692
30	60	0.493	0.055	100%	0.310	0.664
30	100	0.491	0.056	100%	0.351	0.621
30	150	0.487	0.054	98%	0.373	0.596
30	200	0.485	0.051	98%	0.388	0.579
30	300	0.479	0.051	96%	0.401	0.554

Table A.11: *Ceiling: Simulation binary-ordinal: R_h^2 results. Where R_h^2=1 and R_{ht}^2=0.90 and the odds are proportional*

Appendix B: Ordinal-binary

Tables presented here are for an ordinal surrogate and a binary true outcome, for each scenario covered in the simulation for the ordinal-binary setting. The results given are the median R^2 values at the trial and individual level for each scenario, the IQR, the coverage of the confidence intervals and the median 95% confidence interval limits.

No. trials	Trial Size	R_h^2	IQR R_h^2	% Cover CIs	Low CI	Upp CI
5	10	0.452	0.177	100%	0.051	0.999
5	20	0.446	0.145	100%	0.092	0.880
5	40	0.423	0.114	100%	0.137	0.757
5	60	0.410	0.106	100%	0.168	0.687
5	100	0.414	0.096	100%	0.214	0.631
5	150	0.396	0.081	100%	0.233	0.573
5	200	0.412	0.069	99%	0.264	0.567
5	300	0.405	0.068	98%	0.287	0.531
10	10	0.446	0.140	100%	0.054	0.989
10	20	0.430	0.117	100%	0.091	0.863
10	40	0.413	0.072	100%	0.135	0.742
10	60	0.399	0.076	100%	0.164	0.674
10	100	0.395	0.058	100%	0.204	0.610
10	150	0.401	0.063	100%	0.238	0.576
10	200	0.389	0.051	100%	0.248	0.542
10	300	0.397	0.054	100%	0.279	0.522
20	10	0.420	0.104	100%	0.049	0.964
20	20	0.423	0.078	100%	0.089	0.850
20	40	0.405	0.051	100%	0.135	0.730
20	60	0.396	0.058	100%	0.165	0.667
20	100	0.393	0.045	100%	0.205	0.607
20	150	0.393	0.041	100%	0.233	0.567
20	200	0.395	0.039	100%	0.255	0.547
20	300	0.393	0.037	100%	0.277	0.517
30	10	0.424	0.074	100%	0.051	0.964
30	20	0.422	0.062	100%	0.092	0.849
30	40	0.406	0.050	100%	0.135	0.732
30	60	0.397	0.040	100%	0.163	0.669
30	100	0.393	0.033	100%	0.203	0.605
30	150	0.389	0.032	100%	0.231	0.564
30	200	0.391	0.034	100%	0.252	0.542
30	300	0.389	0.033	100%	0.273	0.513

Table B. 1: *Simulation ordinal-binary: Results R_h^2. Where R_h^2=0.64 and R_{ht}^2=0.90 and the odds are proportional*

No. trials	Trial Size	R_{ht}^2	IQR R_h^2	% Cover CIs	Low CI	Upp CI
5	10	0.760	0.376	92%	0.097	0.986
5	20	0.873	0.253	93%	0.269	0.995
5	40	0.876	0.282	91%	0.275	0.995
5	60	0.907	0.165	95%	0.358	0.997
5	100	0.923	0.172	94%	0.408	0.998
5	150	0.928	0.137	94%	0.427	0.998
5	200	0.937	0.109	94%	0.461	0.998
5	300	0.944	0.116	93%	0.489	0.999
10	10	0.588	0.324	55%	0.097	0.913
10	20	0.689	0.295	71%	0.192	0.945
10	40	0.796	0.199	88%	0.338	0.971
10	60	0.823	0.188	92%	0.385	0.976
10	100	0.860	0.143	93%	0.457	0.983
10	150	0.872	0.119	98%	0.485	0.985
10	200	0.883	0.109	97%	0.510	0.987
10	300	0.888	0.104	98%	0.523	0.988
20	10	0.487	0.204	14%	0.133	0.793
20	20	0.640	0.207	37%	0.279	0.877
20	40	0.724	0.156	63%	0.384	0.916
20	60	0.771	0.142	75%	0.452	0.935
20	100	0.820	0.111	89%	0.531	0.953
20	150	0.842	0.106	94%	0.571	0.960
20	200	0.848	0.102	94%	0.583	0.962
20	300	0.862	0.077	97%	0.610	0.967
30	10	0.480	0.197	3%	0.184	0.744
30	20	0.614	0.152	10%	0.317	0.831
30	40	0.718	0.127	38%	0.445	0.889
30	60	0.760	0.103	61%	0.503	0.910
30	100	0.798	0.081	80%	0.560	0.928
30	150	0.825	0.080	88%	0.604	0.940
30	200	0.834	0.088	90%	0.619	0.944
30	300	0.854	0.071	96%	0.653	0.952

Table B. 2: *Simulation ordinal-binary: Results R_{ht}^2. Where R_h^2=0.64 and R_{ht}^2=0.90 and the odds are proportional*

No. trials	Trial Size	R_h^2	IQR R_h^2	% Cover CIs	Low CI	Upp CI
5	10	0.443	0.206	100%	0.046	0.993
5	20	0.426	0.132	100%	0.086	0.853
5	40	0.411	0.098	100%	0.137	0.746
5	60	0.415	0.088	100%	0.173	0.692
5	100	0.409	0.089	100%	0.213	0.626
5	150	0.407	0.082	100%	0.242	0.585
5	200	0.405	0.086	98%	0.261	0.559
5	300	0.403	0.073	96%	0.284	0.529
10	10	0.433	0.138	100%	0.052	0.978
10	20	0.423	0.106	100%	0.087	0.854
10	40	0.406	0.083	100%	0.136	0.737
10	60	0.405	0.072	100%	0.167	0.677
10	100	0.397	0.056	100%	0.203	0.615
10	150	0.395	0.056	100%	0.234	0.569
10	200	0.394	0.050	100%	0.255	0.547
10	300	0.388	0.055	100%	0.272	0.512
20	10	0.429	0.087	100%	0.049	0.970
20	20	0.423	0.075	100%	0.089	0.851
20	40	0.401	0.060	100%	0.130	0.727
20	60	0.400	0.044	100%	0.166	0.673
20	100	0.389	0.047	100%	0.200	0.602
20	150	0.389	0.038	100%	0.230	0.564
20	200	0.390	0.042	100%	0.250	0.542
20	300	0.389	0.038	100%	0.273	0.512
30	10	0.425	0.079	100%	0.048	0.963
30	20	0.419	0.059	100%	0.090	0.845
30	40	0.395	0.042	100%	0.129	0.718
30	60	0.395	0.042	100%	0.162	0.666
30	100	0.393	0.034	100%	0.203	0.605
30	150	0.389	0.034	100%	0.231	0.565
30	200	0.390	0.032	100%	0.250	0.542
30	300	0.384	0.031	100%	0.270	0.507

Table B. 3: *Simulation ordinal-binary: Results R_h^2. Where R_h^2=0.64 and R_{ht}^2=0.30 and the odds are proportional*

No. trials	Trial Size	R^2_{ht}	IQR R^2_h	% Cover CIs	Low CI	Upp CI
5	10	0.548	0.506	84%	0.010	0.956
5	20	0.645	0.487	77%	0.029	0.972
5	40	0.652	0.445	80%	0.031	0.973
5	60	0.639	0.457	78%	0.026	0.971
5	100	0.643	0.434	85%	0.028	0.972
5	150	0.677	0.465	80%	0.041	0.977
5	200	0.658	0.489	77%	0.033	0.974
5	300	0.668	0.429	80%	0.037	0.976
10	10	0.324	0.406	97%	0.005	0.788
10	20	0.409	0.342	95%	0.015	0.836
10	40	0.367	0.363	95%	0.009	0.814
10	60	0.416	0.338	97%	0.017	0.840
10	100	0.421	0.357	94%	0.018	0.842
10	150	0.396	0.363	95%	0.013	0.830
10	200	0.385	0.291	93%	0.011	0.823
10	300	0.418	0.376	94%	0.017	0.841
20	10	0.285	0.266	96%	0.020	0.645
20	20	0.298	0.278	95%	0.025	0.656
20	40	0.336	0.249	96%	0.040	0.688
20	60	0.321	0.248	95%	0.034	0.676
20	100	0.330	0.222	95%	0.037	0.682
20	150	0.337	0.235	96%	0.040	0.688
20	200	0.352	0.220	97%	0.048	0.700
20	300	0.344	0.241	96%	0.044	0.694
30	10	0.229	0.199	92%	0.023	0.530
30	20	0.279	0.180	96%	0.045	0.579
30	40	0.289	0.216	94%	0.050	0.588
30	60	0.324	0.202	96%	0.069	0.620
30	100	0.303	0.180	94%	0.057	0.601
30	150	0.324	0.190	97%	0.069	0.620
30	200	0.317	0.216	97%	0.065	0.614
30	300	0.323	0.199	96%	0.068	0.619

Table B. 4: *Simulation ordinal-binary: Results R^2_{ht}. Where R^2_h=0.64 and R^2_{ht}=0.30 and the odds are proportional*

No. trials	Trial Size	R_h^2	IQR R_h^2	% Cover CIs	Low CI	Upp CI
5	10	0.274	0.173	100%	0.018	0.856
5	20	0.218	0.127	100%	0.017	0.648
5	40	0.204	0.079	100%	0.031	0.512
5	60	0.190	0.071	100%	0.038	0.443
5	100	0.181	0.063	100%	0.054	0.373
5	150	0.185	0.050	100%	0.071	0.338
5	200	0.176	0.052	100%	0.078	0.309
5	300	0.181	0.042	100%	0.096	0.288
10	10	0.294	0.135	100%	0.021	0.871
10	20	0.221	0.084	100%	0.020	0.644
10	40	0.199	0.057	100%	0.030	0.506
10	60	0.187	0.050	100%	0.039	0.431
10	100	0.185	0.039	100%	0.056	0.372
10	150	0.182	0.035	100%	0.070	0.333
10	200	0.178	0.034	100%	0.078	0.309
10	300	0.175	0.029	100%	0.092	0.280
20	10	0.272	0.088	100%	0.022	0.848
20	20	0.222	0.059	100%	0.023	0.648
20	40	0.196	0.043	100%	0.031	0.501
20	60	0.186	0.036	100%	0.040	0.430
20	100	0.178	0.028	100%	0.054	0.367
20	150	0.175	0.021	100%	0.067	0.326
20	200	0.175	0.023	100%	0.077	0.304
20	300	0.176	0.021	100%	0.092	0.280
30	10	0.275	0.074	100%	0.021	0.854
30	20	0.223	0.047	100%	0.024	0.650
30	40	0.199	0.033	100%	0.032	0.502
30	60	0.187	0.032	100%	0.040	0.432
30	100	0.182	0.022	100%	0.056	0.369
30	150	0.177	0.019	100%	0.068	0.328
30	200	0.176	0.019	100%	0.078	0.306
30	300	0.174	0.018	100%	0.091	0.278

Table B. 5: *Simulation ordinal-binary: Results R_h^2. Where R_h^2=0.30 and R_{ht}^2=0.90 and the odds are proportional*

No. trials	Trial Size	R_{ht}^2	IQR R_h^2	% Cover CIs	Low CI	Upp CI
5	10	0.703	0.335	89%	0.054	0.980
5	20	0.814	0.295	92%	0.163	0.991
5	40	0.847	0.250	92%	0.216	0.994
5	60	0.895	0.187	94%	0.323	0.997
5	100	0.920	0.171	98%	0.400	0.998
5	150	0.922	0.122	95%	0.404	0.998
5	200	0.927	0.135	91%	0.422	0.998
5	300	0.947	0.112	95%	0.503	0.999
10	10	0.504	0.367	44%	0.047	0.881
10	20	0.619	0.304	61%	0.123	0.923
10	40	0.771	0.211	85%	0.297	0.965
10	60	0.789	0.195	86%	0.325	0.969
10	100	0.837	0.154	94%	0.411	0.979
10	150	0.854	0.133	95%	0.446	0.982
10	200	0.861	0.119	97%	0.459	0.983
10	300	0.881	0.101	98%	0.506	0.987
20	10	0.436	0.235	4%	0.094	0.761
20	20	0.566	0.218	20%	0.202	0.839
20	40	0.695	0.158	52%	0.345	0.903
20	60	0.746	0.129	71%	0.415	0.925
20	100	0.790	0.126	85%	0.483	0.942
20	150	0.816	0.119	89%	0.525	0.951
20	200	0.838	0.108	92%	0.563	0.959
20	300	0.864	0.084	95%	0.613	0.967
30	10	0.388	0.221	1%	0.111	0.675
30	20	0.552	0.187	3%	0.252	0.793
30	40	0.675	0.126	22%	0.390	0.866
30	60	0.725	0.118	46%	0.454	0.893
30	100	0.790	0.102	72%	0.548	0.924
30	150	0.810	0.101	82%	0.580	0.934
30	200	0.826	0.079	87%	0.605	0.941
30	300	0.851	0.079	92%	0.648	0.951

Table B. 6: *Simulation ordinal-binary: Results R_{ht}^2. Where R_h^2=0.30 and R_{ht}^2=0.90 and the odds are proportional*

No. trials	Trial Size	R_h^2	IQR R_h^2	% Cover CIs	Low CI	Upp CI
5	10	0.296	0.184	100%	0.017	0.868
5	20	0.228	0.107	100%	0.017	0.662
5	40	0.195	0.095	100%	0.026	0.507
5	60	0.195	0.069	100%	0.040	0.445
5	100	0.182	0.055	100%	0.054	0.375
5	150	0.180	0.050	100%	0.068	0.333
5	200	0.180	0.045	100%	0.080	0.312
5	300	0.181	0.038	100%	0.095	0.287
10	10	0.266	0.115	100%	0.019	0.851
10	20	0.228	0.091	100%	0.022	0.658
10	40	0.196	0.056	100%	0.030	0.501
10	60	0.187	0.048	100%	0.038	0.434
10	100	0.180	0.039	100%	0.054	0.368
10	150	0.179	0.030	100%	0.069	0.332
10	200	0.178	0.032	100%	0.079	0.307
10	300	0.173	0.030	100%	0.090	0.278
20	10	0.279	0.081	100%	0.023	0.855
20	20	0.223	0.061	100%	0.024	0.650
20	40	0.193	0.040	100%	0.031	0.498
20	60	0.182	0.030	100%	0.038	0.429
20	100	0.177	0.033	100%	0.053	0.362
20	150	0.175	0.026	100%	0.067	0.325
20	200	0.173	0.021	100%	0.077	0.302
20	300	0.175	0.019	100%	0.092	0.280
30	10	0.280	0.067	100%	0.022	0.854
30	20	0.221	0.045	100%	0.024	0.644
30	40	0.193	0.034	100%	0.031	0.495
30	60	0.184	0.026	100%	0.039	0.428
30	100	0.179	0.024	100%	0.054	0.366
30	150	0.175	0.021	100%	0.067	0.325
30	200	0.176	0.016	100%	0.078	0.305
30	300	0.172	0.017	100%	0.089	0.275

Table B. 7: *Simulation ordinal-binary: Results R_h^2. Where R_h^2=0.30 and R_{ht}^2=0.30 and the odds are proportional*

No. trials	Trial Size	R^2_{ht}	IQR R^2_h	% Cover CIs	Low CI	Upp CI
5	10	0.619	0.431	84%	0.021	0.968
5	20	0.676	0.463	80%	0.040	0.977
5	40	0.590	0.471	84%	0.015	0.964
5	60	0.596	0.503	80%	0.016	0.965
5	100	0.672	0.506	83%	0.038	0.976
5	150	0.613	0.439	85%	0.020	0.967
5	200	0.665	0.451	80%	0.036	0.975
5	300	0.666	0.391	80%	0.036	0.975
10	10	0.285	0.350	96%	0.000	0.763
10	20	0.350	0.284	96%	0.007	0.804
10	40	0.383	0.373	97%	0.011	0.823
10	60	0.385	0.353	98%	0.011	0.823
10	100	0.399	0.357	95%	0.013	0.831
10	150	0.454	0.356	93%	0.027	0.858
10	200	0.372	0.351	96%	0.009	0.816
10	300	0.424	0.326	97%	0.019	0.844
20	10	0.225	0.232	94%	0.007	0.589
20	20	0.274	0.207	97%	0.017	0.635
20	40	0.301	0.228	96%	0.026	0.659
20	60	0.314	0.198	96%	0.030	0.669
20	100	0.324	0.271	98%	0.035	0.678
20	150	0.337	0.242	97%	0.041	0.688
20	200	0.342	0.248	97%	0.043	0.692
20	300	0.284	0.249	96%	0.020	0.644
30	10	0.196	0.155	90%	0.012	0.495
30	20	0.235	0.175	92%	0.025	0.535
30	40	0.302	0.214	94%	0.057	0.601
30	60	0.285	0.173	95%	0.048	0.584
30	100	0.297	0.199	98%	0.054	0.596
30	150	0.321	0.216	95%	0.068	0.618
30	200	0.307	0.190	96%	0.060	0.605
30	300	0.321	0.190	96%	0.068	0.618

Table B. 8: *Simulation ordinal-binary: Results R^2_{ht}. Where R^2_h=0.30 and R^2_{ht}=0.30 and the odds are proportional*

No. trials	Trial Size	R_h^2	IQR R_h^2	% Cover CIs	Low CI	Upp CI
5	10	0.449	0.204	100%	0.051	0.996
5	20	0.435	0.149	100%	0.087	0.862
5	40	0.412	0.128	100%	0.133	0.744
5	60	0.416	0.097	100%	0.172	0.695
5	100	0.406	0.081	100%	0.210	0.622
5	150	0.402	0.088	99%	0.240	0.581
5	200	0.402	0.076	100%	0.259	0.558
5	300	0.399	0.070	97%	0.281	0.526
10	10	0.437	0.158	100%	0.054	0.982
10	20	0.416	0.106	100%	0.086	0.849
10	40	0.403	0.083	100%	0.130	0.728
10	60	0.398	0.067	100%	0.163	0.672
10	100	0.403	0.060	100%	0.210	0.615
10	150	0.393	0.059	100%	0.233	0.569
10	200	0.387	0.045	100%	0.248	0.541
10	300	0.391	0.048	100%	0.275	0.516
20	10	0.414	0.102	100%	0.049	0.953
20	20	0.414	0.074	100%	0.088	0.842
20	40	0.400	0.053	100%	0.131	0.728
20	60	0.392	0.052	100%	0.160	0.665
20	100	0.390	0.044	100%	0.200	0.603
20	150	0.388	0.040	100%	0.228	0.563
20	200	0.387	0.038	100%	0.248	0.538
20	300	0.383	0.043	100%	0.268	0.506
30	10	0.423	0.076	100%	0.051	0.963
30	20	0.417	0.056	100%	0.091	0.844
30	40	0.398	0.043	100%	0.130	0.721
30	60	0.393	0.038	100%	0.162	0.665
30	100	0.386	0.036	100%	0.197	0.599
30	150	0.387	0.034	100%	0.228	0.561
30	200	0.386	0.032	100%	0.247	0.538
30	300	0.382	0.028	100%	0.267	0.505

Table B. 9*: Simulation ordinal-binary: Results R_h^2. Where R_h^2=0.64 and R_{ht}^2=0.90 and the odds are non-proportional*

No. trials	Trial Size	R^2_{ht}	IQR R^2_h	% Cover CIs	Low CI	Upp CI
5	10	0.758	0.401	85%	0.096	0.986
5	20	0.871	0.260	90%	0.264	0.995
5	40	0.880	0.206	94%	0.285	0.996
5	60	0.905	0.185	94%	0.350	0.997
5	100	0.925	0.149	96%	0.416	0.998
5	150	0.930	0.141	95%	0.432	0.998
5	200	0.933	0.128	95%	0.447	0.998
5	300	0.931	0.111	96%	0.436	0.998
10	10	0.562	0.299	52%	0.080	0.908
10	20	0.693	0.267	74%	0.196	0.946
10	40	0.770	0.214	86%	0.296	0.965
10	60	0.808	0.202	90%	0.358	0.973
10	100	0.828	0.150	94%	0.394	0.977
10	150	0.860	0.133	94%	0.458	0.983
10	200	0.879	0.118	98%	0.501	0.986
10	300	0.885	0.099	99%	0.515	0.987
20	10	0.505	0.196	11%	0.146	0.805
20	20	0.616	0.213	33%	0.252	0.866
20	40	0.717	0.155	59%	0.374	0.913
20	60	0.756	0.156	70%	0.430	0.929
20	100	0.805	0.131	83%	0.507	0.948
20	150	0.830	0.116	88%	0.550	0.956
20	200	0.839	0.114	93%	0.566	0.959
20	300	0.859	0.088	96%	0.603	0.966
30	10	0.495	0.195	2%	0.195	0.754
30	20	0.615	0.148	10%	0.318	0.831
30	40	0.697	0.138	36%	0.418	0.878
30	60	0.756	0.112	56%	0.498	0.908
30	100	0.797	0.097	76%	0.560	0.928
30	150	0.819	0.089	85%	0.594	0.937
30	200	0.834	0.083	90%	0.618	0.944
30	300	0.852	0.074	91%	0.649	0.951

Table B. 10*: Simulation ordinal-binary: Results R^2_{ht}. Where R^2_h=0.64 and R^2_{ht}=0.90 and the odds are non-proportional*

No. trials	Trial Size	R_h^2	IQR R_h^2	% Cover CIs	Low CI	Upp CI
5	10	0.620	0.220	100%	0.107	1.000
5	20	0.709	0.177	100%	0.277	1.000
5	40	0.719	0.161	99%	0.392	0.998
5	60	0.733	0.140	97%	0.461	0.969
5	100	0.726	0.136	96%	0.516	0.916
5	150	0.740	0.136	91%	0.568	0.895
5	200	0.753	0.124	88%	0.604	0.889
5	300	0.733	0.135	75%	0.611	0.847
10	10	0.615	0.133	100%	0.113	1.000
10	20	0.689	0.123	100%	0.255	1.000
10	40	0.710	0.091	100%	0.391	0.984
10	60	0.723	0.098	100%	0.460	0.955
10	100	0.718	0.088	99%	0.510	0.904
10	150	0.717	0.091	98%	0.549	0.872
10	200	0.714	0.084	96%	0.569	0.848
10	300	0.717	0.096	90%	0.596	0.828
20	10	0.603	0.093	100%	0.108	1.000
20	20	0.669	0.077	100%	0.252	1.000
20	40	0.693	0.084	100%	0.380	0.969
20	60	0.704	0.066	100%	0.441	0.937
20	100	0.707	0.071	100%	0.502	0.891
20	150	0.711	0.069	100%	0.542	0.865
20	200	0.710	0.065	100%	0.563	0.846
20	300	0.707	0.071	100%	0.586	0.819
30	10	0.595	0.078	100%	0.107	1.000
30	20	0.666	0.077	100%	0.247	1.000
30	40	0.697	0.057	100%	0.379	0.971
30	60	0.697	0.053	100%	0.436	0.932
30	100	0.701	0.049	100%	0.496	0.887
30	150	0.698	0.062	100%	0.532	0.855
30	200	0.707	0.052	100%	0.562	0.842
30	300	0.704	0.054	100%	0.587	0.816

Table B. 11: *Ceiling: Simulation ordinal-binary: Results R_h^2. Where R_h^2=1 and R_{ht}^2=0.90 and the odds are proportional*

Appendix C: Ordinal-ordinal

Tables presented here are for an ordinal surrogate and an ordinal true outcome, for each scenario covered in the simulation for the ordinal-ordinal setting. The results given are the median R^2 values at the trial and individual level for each scenario, the IQR, the coverage of the confidence intervals and the median 95% confidence interval limits.

No. trials	Trial Size	R_h^2	IQR R_h^2	% Cover CIs	Low CI	Upp CI
5	10	0.562	0.175	100%	0.147	0.893
5	20	0.535	0.095	100%	0.190	0.819
5	40	0.538	0.072	100%	0.280	0.760
5	60	0.536	0.050	100%	0.322	0.724
5	100	0.536	0.045	100%	0.370	0.687
5	150	0.533	0.038	100%	0.397	0.659
5	200	0.539	0.034	100%	0.421	0.648
5	300	0.538	0.035	100%	0.442	0.629
10	10	0.538	0.101	100%	0.133	0.877
10	20	0.532	0.066	100%	0.198	0.818
10	40	0.527	0.050	100%	0.272	0.751
10	60	0.534	0.037	100%	0.322	0.722
10	100	0.533	0.033	100%	0.367	0.683
10	150	0.534	0.025	100%	0.399	0.659
10	200	0.533	0.027	100%	0.416	0.643
10	300	0.532	0.022	100%	0.436	0.622
20	10	0.539	0.069	100%	0.143	0.880
20	20	0.528	0.054	100%	0.194	0.814
20	40	0.531	0.036	100%	0.277	0.752
20	60	0.530	0.029	100%	0.318	0.717
20	100	0.532	0.023	100%	0.366	0.682
20	150	0.529	0.019	100%	0.393	0.654
20	200	0.531	0.017	100%	0.413	0.640
20	300	0.530	0.016	100%	0.433	0.620
30	10	0.549	0.055	100%	0.146	0.881
30	20	0.526	0.040	100%	0.194	0.812
30	40	0.529	0.028	100%	0.274	0.750
30	60	0.530	0.021	100%	0.318	0.718
30	100	0.528	0.021	100%	0.363	0.679
30	150	0.528	0.015	100%	0.393	0.654
30	200	0.529	0.015	100%	0.411	0.638
30	300	0.529	0.013	100%	0.433	0.619

Table C. 1 *Simulation ordinal-ordinal: Results R_h^2. Where R_h^2=0.64 and R_{ht}^2=0.90 and the odds are proportional*

No. trials	Trial Size	R^2_{ht}	IQR R^2_h	% Cover CIs	Low CI	Upp CI
5	10	0.853	0.228	95%	0.228	0.994
5	20	0.921	0.154	97%	0.403	0.998
5	40	0.943	0.116	95%	0.485	0.999
5	60	0.948	0.085	95%	0.507	0.999
5	100	0.959	0.077	92%	0.563	0.999
5	150	0.960	0.092	94%	0.571	0.999
5	200	0.955	0.080	94%	0.540	0.999
5	300	0.955	0.079	95%	0.542	0.999
10	10	0.756	0.213	85%	0.274	0.962
10	20	0.824	0.160	95%	0.386	0.977
10	40	0.875	0.118	99%	0.491	0.986
10	60	0.892	0.085	99%	0.534	0.989
10	100	0.902	0.087	100%	0.559	0.990
10	150	0.912	0.101	98%	0.585	0.991
10	200	0.918	0.089	98%	0.604	0.992
10	300	0.916	0.080	98%	0.597	0.992
20	10	0.682	0.166	48%	0.329	0.897
20	20	0.793	0.099	86%	0.487	0.943
20	40	0.857	0.083	97%	0.600	0.965
20	60	0.877	0.070	98%	0.639	0.972
20	100	0.887	0.071	99%	0.659	0.974
20	150	0.894	0.078	99%	0.676	0.977
20	200	0.895	0.072	98%	0.676	0.977
20	300	0.899	0.066	99%	0.687	0.978
30	10	0.683	0.114	25%	0.400	0.872
30	20	0.792	0.086	77%	0.549	0.926
30	40	0.847	0.078	92%	0.641	0.950
30	60	0.869	0.073	96%	0.680	0.958
30	100	0.880	0.058	98%	0.700	0.963
30	150	0.884	0.062	96%	0.709	0.964
30	200	0.891	0.053	98%	0.721	0.967
30	300	0.895	0.047	99%	0.730	0.969

Table C. 2 *Simulation ordinal-ordinal: Results R^2_{ht}. Where R^2_h=0.64 and R^2_{ht}=0.90 and the odds are proportional*

No. trials	Trial Size	R_h^2	IQR R_h^2	% Cover CIs	Low CI	Upp CI
5	10	0.537	0.153	100%	0.131	0.886
5	20	0.520	0.105	100%	0.186	0.814
5	40	0.527	0.071	100%	0.272	0.750
5	60	0.534	0.059	100%	0.320	0.723
5	100	0.535	0.042	100%	0.367	0.685
5	150	0.533	0.037	100%	0.396	0.659
5	200	0.533	0.035	100%	0.415	0.642
5	300	0.534	0.031	100%	0.437	0.625
10	10	0.538	0.102	100%	0.135	0.878
10	20	0.533	0.075	100%	0.194	0.816
10	40	0.527	0.048	100%	0.270	0.750
10	60	0.526	0.038	100%	0.315	0.715
10	100	0.525	0.033	100%	0.359	0.676
10	150	0.527	0.026	100%	0.391	0.654
10	200	0.527	0.026	100%	0.409	0.637
10	300	0.528	0.025	100%	0.432	0.618
20	10	0.539	0.076	100%	0.143	0.874
20	20	0.525	0.047	100%	0.195	0.810
20	40	0.526	0.036	100%	0.271	0.747
20	60	0.524	0.032	100%	0.313	0.713
20	100	0.525	0.022	100%	0.359	0.676
20	150	0.526	0.021	100%	0.391	0.651
20	200	0.526	0.019	100%	0.409	0.636
20	300	0.524	0.017	100%	0.428	0.615
30	10	0.543	0.057	100%	0.143	0.879
30	20	0.523	0.039	100%	0.193	0.811
30	40	0.526	0.030	100%	0.272	0.748
30	60	0.526	0.026	100%	0.314	0.714
30	100	0.523	0.020	100%	0.358	0.674
30	150	0.525	0.016	100%	0.389	0.651
30	200	0.526	0.015	100%	0.408	0.635
30	300	0.524	0.013	100%	0.427	0.614

Table C.3: *Simulation ordinal-ordinal: Results R_h^2. Where R_h^2=0.64 and R_{ht}^2=0.30 and the odds are proportional*

No. trials	Trial Size	R^2_{ht}	IQR R^2_h	% Cover CIs	Low CI	Upp CI
5	10	0.699	0.394	78%	0.052	0.980
5	20	0.680	0.425	77%	0.042	0.977
5	40	0.673	0.431	80%	0.039	0.976
5	60	0.645	0.502	82%	0.028	0.972
5	100	0.629	0.471	81%	0.024	0.970
5	150	0.660	0.430	80%	0.034	0.974
5	200	0.706	0.415	78%	0.056	0.980
5	300	0.689	0.404	81%	0.046	0.978
10	10	0.443	0.380	93%	0.024	0.853
10	20	0.428	0.389	95%	0.020	0.846
10	40	0.449	0.329	94%	0.026	0.856
10	60	0.448	0.324	95%	0.025	0.855
10	100	0.434	0.324	93%	0.021	0.849
10	150	0.446	0.345	95%	0.025	0.855
10	200	0.425	0.313	96%	0.019	0.844
10	300	0.425	0.350	91%	0.019	0.844
20	10	0.345	0.279	96%	0.042	0.697
20	20	0.343	0.256	95%	0.043	0.693
20	40	0.368	0.235	95%	0.056	0.712
20	60	0.348	0.238	97%	0.046	0.697
20	100	0.357	0.213	96%	0.050	0.704
20	150	0.340	0.247	98%	0.042	0.690
20	200	0.370	0.228	97%	0.057	0.713
20	300	0.372	0.240	98%	0.058	0.715
30	10	0.371	0.188	95%	0.099	0.661
30	20	0.344	0.199	93%	0.082	0.638
30	40	0.347	0.198	96%	0.083	0.640
30	60	0.354	0.182	95%	0.088	0.646
30	100	0.348	0.193	96%	0.084	0.641
30	150	0.319	0.185	98%	0.066	0.616
30	200	0.341	0.195	97%	0.079	0.634
30	300	0.340	0.199	96%	0.079	0.634

Table C.4: *Simulation ordinal-ordinal: Results R^2_{ht}. Where R^2_h=0.64 and R^2_{ht}=0.30 and the odds are proportional*

No. trials	Trial Size	R_h^2	IQR R_h^2	% Cover CIs	Low CI	Upp CI
5	10	0.301	0.142	100%	0.029	0.738
5	20	0.264	0.112	100%	0.040	0.604
5	40	0.256	0.075	100%	0.064	0.513
5	60	0.251	0.065	100%	0.083	0.462
5	100	0.248	0.048	100%	0.110	0.413
5	150	0.244	0.039	100%	0.128	0.379
5	200	0.243	0.030	100%	0.141	0.360
5	300	0.245	0.028	100%	0.160	0.339
10	10	0.308	0.102	100%	0.037	0.732
10	20	0.257	0.085	100%	0.039	0.597
10	40	0.246	0.050	100%	0.062	0.501
10	60	0.246	0.044	100%	0.083	0.456
10	100	0.247	0.036	100%	0.110	0.412
10	150	0.242	0.026	100%	0.127	0.376
10	200	0.242	0.021	100%	0.139	0.356
10	300	0.240	0.019	100%	0.155	0.334
20	10	0.306	0.080	100%	0.040	0.734
20	20	0.266	0.047	100%	0.045	0.603
20	40	0.253	0.034	100%	0.065	0.506
20	60	0.247	0.029	100%	0.083	0.456
20	100	0.244	0.023	100%	0.108	0.407
20	150	0.241	0.020	100%	0.126	0.375
20	200	0.241	0.017	100%	0.140	0.356
20	300	0.239	0.014	100%	0.155	0.334
30	10	0.307	0.065	100%	0.042	0.734
30	20	0.269	0.040	100%	0.046	0.606
30	40	0.248	0.030	100%	0.062	0.503
30	60	0.244	0.024	100%	0.081	0.453
30	100	0.241	0.018	100%	0.106	0.404
30	150	0.240	0.015	100%	0.126	0.374
30	200	0.241	0.014	100%	0.139	0.356
30	300	0.240	0.011	100%	0.155	0.334

Table C.5: *Simulation ordinal-ordinal: Results R_h^2. Where R_h^2=0.30 and R_{ht}^2=0.90 and the odds are proportional*

No. trials	Trial Size	R_{ht}^2	IQR R_h^2	% Cover CIs	Low CI	Upp CI
5	10	0.795	0.313	89%	0.137	0.990
5	20	0.880	0.243	94%	0.286	0.996
5	40	0.923	0.155	96%	0.410	0.998
5	60	0.930	0.126	96%	0.433	0.998
5	100	0.929	0.119	95%	0.429	0.998
5	150	0.949	0.097	94%	0.515	0.999
5	200	0.951	0.091	96%	0.522	0.999
5	300	0.956	0.077	96%	0.546	0.999
10	10	0.634	0.272	63%	0.136	0.928
10	20	0.737	0.198	86%	0.249	0.957
10	40	0.850	0.162	94%	0.436	0.981
10	60	0.871	0.119	98%	0.482	0.985
10	100	0.894	0.089	98%	0.538	0.989
10	150	0.896	0.091	98%	0.542	0.989
10	200	0.911	0.092	99%	0.585	0.991
10	300	0.906	0.095	99%	0.570	0.991
20	10	0.568	0.218	22%	0.205	0.842
20	20	0.713	0.151	60%	0.367	0.911
20	40	0.805	0.109	88%	0.505	0.948
20	60	0.844	0.078	96%	0.575	0.961
20	100	0.874	0.075	99%	0.632	0.970
20	150	0.885	0.076	97%	0.655	0.974
20	200	0.885	0.075	97%	0.656	0.974
20	300	0.894	0.069	100%	0.676	0.977
30	10	0.542	0.181	4%	0.242	0.788
30	20	0.706	0.127	34%	0.427	0.883
30	40	0.801	0.100	75%	0.566	0.930
30	60	0.833	0.073	92%	0.618	0.944
30	100	0.867	0.071	97%	0.677	0.958
30	150	0.872	0.062	96%	0.686	0.960
30	200	0.877	0.053	96%	0.694	0.961
30	300	0.885	0.051	98%	0.711	0.965

Table C.6: *Simulation ordinal-ordinal: Results R_{ht}^2. Where R_h^2=0.30 and R_{ht}^2=0.90 and the odds are proportional*

No. trials	Trial Size	R_h^2	IQR R_h^2	% Cover CIs	Low CI	Upp CI
5	10	0.293	0.154	100%	0.023	0.736
5	20	0.267	0.110	100%	0.040	0.613
5	40	0.249	0.076	100%	0.062	0.504
5	60	0.236	0.051	100%	0.076	0.447
5	100	0.240	0.048	100%	0.104	0.404
5	150	0.243	0.040	100%	0.128	0.376
5	200	0.242	0.034	100%	0.139	0.359
5	300	0.241	0.027	100%	0.156	0.336
10	10	0.306	0.100	100%	0.037	0.736
10	20	0.272	0.085	100%	0.045	0.613
10	40	0.246	0.054	100%	0.062	0.499
10	60	0.244	0.043	100%	0.080	0.455
10	100	0.245	0.036	100%	0.109	0.408
10	150	0.241	0.033	100%	0.126	0.375
10	200	0.240	0.023	100%	0.138	0.355
10	300	0.241	0.019	100%	0.156	0.335
20	10	0.302	0.068	100%	0.039	0.732
20	20	0.268	0.042	100%	0.042	0.604
20	40	0.249	0.033	100%	0.064	0.501
20	60	0.243	0.026	100%	0.080	0.453
20	100	0.239	0.023	100%	0.105	0.403
20	150	0.238	0.018	100%	0.125	0.371
20	200	0.237	0.016	100%	0.137	0.352
20	300	0.238	0.013	100%	0.153	0.331
30	10	0.308	0.057	100%	0.042	0.734
30	20	0.264	0.042	100%	0.043	0.606
30	40	0.250	0.027	100%	0.064	0.503
30	60	0.245	0.022	100%	0.082	0.453
30	100	0.241	0.019	100%	0.106	0.404
30	150	0.241	0.016	100%	0.126	0.374
30	200	0.239	0.013	100%	0.138	0.354
30	300	0.239	0.012	100%	0.154	0.333

Table C.7: *Simulation ordinal-ordinal: Results R_h^2. Where R_h^2=0.30 and R_{ht}^2=0.30 and the odds are proportional*

No. trials	Trial Size	R^2_{ht}	IQR R^2_h	% Cover CIs	Low CI	Upp CI
5	10	0.651	0.431	82%	0.030	0.973
5	20	0.650	0.450	82%	0.030	0.973
5	40	0.672	0.469	81%	0.038	0.976
5	60	0.609	0.435	82%	0.019	0.967
5	100	0.680	0.442	80%	0.042	0.977
5	150	0.694	0.483	74%	0.049	0.979
5	200	0.678	0.423	80%	0.041	0.977
5	300	0.678	0.451	78%	0.041	0.977
10	10	0.372	0.352	94%	0.009	0.819
10	20	0.371	0.397	95%	0.009	0.816
10	40	0.398	0.349	96%	0.013	0.831
10	60	0.424	0.353	94%	0.019	0.844
10	100	0.412	0.336	96%	0.016	0.838
10	150	0.427	0.320	94%	0.019	0.845
10	200	0.408	0.333	96%	0.015	0.836
10	300	0.452	0.334	95%	0.026	0.857
20	10	0.287	0.255	95%	0.021	0.647
20	20	0.345	0.248	98%	0.044	0.694
20	40	0.374	0.218	94%	0.059	0.716
20	60	0.346	0.226	97%	0.045	0.695
20	100	0.340	0.238	96%	0.042	0.691
20	150	0.324	0.232	96%	0.035	0.677
20	200	0.351	0.252	93%	0.047	0.699
20	300	0.348	0.241	95%	0.045	0.697
30	10	0.265	0.192	95%	0.038	0.567
30	20	0.293	0.216	95%	0.052	0.592
30	40	0.306	0.187	97%	0.059	0.604
30	60	0.328	0.186	97%	0.071	0.623
30	100	0.319	0.186	96%	0.067	0.616
30	150	0.329	0.170	99%	0.072	0.624
30	200	0.342	0.179	97%	0.080	0.635
30	300	0.317	0.209	96%	0.065	0.613

Table C.8: *Simulation ordinal-ordinal: Results R^2_{ht}. Where R^2_h=0.30 and R^2_{ht}=0.30 and the odds are proportional*

No. trials	Trial Size	R_h^2	IQR R_h^2	% Cover CIs	Low CI	Upp CI
5	10	0.515	0.155	100%	0.128	0.873
5	20	0.517	0.112	100%	0.181	0.811
5	40	0.530	0.080	100%	0.274	0.753
5	60	0.518	0.067	100%	0.304	0.710
5	100	0.526	0.042	100%	0.359	0.678
5	150	0.524	0.037	100%	0.387	0.650
5	200	0.521	0.039	100%	0.402	0.631
5	300	0.520	0.032	100%	0.422	0.611
10	10	0.528	0.097	100%	0.131	0.873
10	20	0.521	0.069	100%	0.190	0.810
10	40	0.515	0.055	100%	0.260	0.741
10	60	0.514	0.042	100%	0.301	0.706
10	100	0.511	0.038	100%	0.344	0.665
10	150	0.516	0.029	100%	0.380	0.642
10	200	0.513	0.027	100%	0.394	0.625
10	300	0.516	0.022	100%	0.420	0.607
20	10	0.535	0.085	100%	0.136	0.877
20	20	0.517	0.051	100%	0.187	0.806
20	40	0.513	0.037	100%	0.259	0.738
20	60	0.514	0.028	100%	0.303	0.705
20	100	0.512	0.027	100%	0.347	0.664
20	150	0.512	0.021	100%	0.376	0.638
20	200	0.511	0.022	100%	0.393	0.622
20	300	0.510	0.018	100%	0.414	0.601
30	10	0.533	0.057	100%	0.141	0.874
30	20	0.511	0.036	100%	0.183	0.800
30	40	0.512	0.029	100%	0.259	0.738
30	60	0.510	0.025	100%	0.300	0.701
30	100	0.510	0.020	100%	0.345	0.663
30	150	0.510	0.020	100%	0.374	0.636
30	200	0.510	0.018	100%	0.392	0.621
30	300	0.510	0.013	100%	0.414	0.602

Table C.9*: Simulation ordinal-ordinal: Results R_h^2. Where R_h^2=0.64 and R_{ht}^2=0.90 and the odds are non-proportional*

No. trials	Trial Size	R^2_{ht}	IQR R^2_h	% Cover CIs	Low CI	Upp CI
5	10	0.850	0.268	90%	0.223	0.994
5	20	0.917	0.146	95%	0.388	0.998
5	40	0.931	0.121	94%	0.437	0.998
5	60	0.942	0.127	93%	0.483	0.999
5	100	0.936	0.128	95%	0.458	0.998
5	150	0.948	0.105	94%	0.510	0.999
5	200	0.949	0.093	95%	0.514	0.999
5	300	0.943	0.099	94%	0.487	0.999
10	10	0.710	0.180	83%	0.216	0.950
10	20	0.790	0.173	92%	0.327	0.970
10	40	0.862	0.126	97%	0.462	0.984
10	60	0.883	0.127	99%	0.512	0.987
10	100	0.891	0.113	99%	0.529	0.988
10	150	0.892	0.098	98%	0.532	0.988
10	200	0.907	0.088	100%	0.572	0.991
10	300	0.910	0.082	97%	0.579	0.991
20	10	0.681	0.164	48%	0.327	0.897
20	20	0.785	0.127	82%	0.474	0.941
20	40	0.842	0.084	95%	0.571	0.960
20	60	0.860	0.084	98%	0.606	0.966
20	100	0.875	0.074	98%	0.636	0.971
20	150	0.885	0.077	98%	0.656	0.974
20	200	0.888	0.063	98%	0.662	0.975
20	300	0.895	0.063	100%	0.676	0.977
30	10	0.673	0.134	24%	0.387	0.866
30	20	0.771	0.105	66%	0.519	0.915
30	40	0.828	0.074	92%	0.610	0.942
30	60	0.857	0.060	95%	0.658	0.953
30	100	0.872	0.064	99%	0.685	0.960
30	150	0.871	0.062	96%	0.683	0.959
30	200	0.876	0.055	95%	0.693	0.961
30	300	0.881	0.054	97%	0.703	0.963

Table C.10*: Simulation ordinal-ordinal: Results R^2_{ht}. Where R^2_h=0.64 and R^2_{ht}=0.90 and the odds are non-proportional*

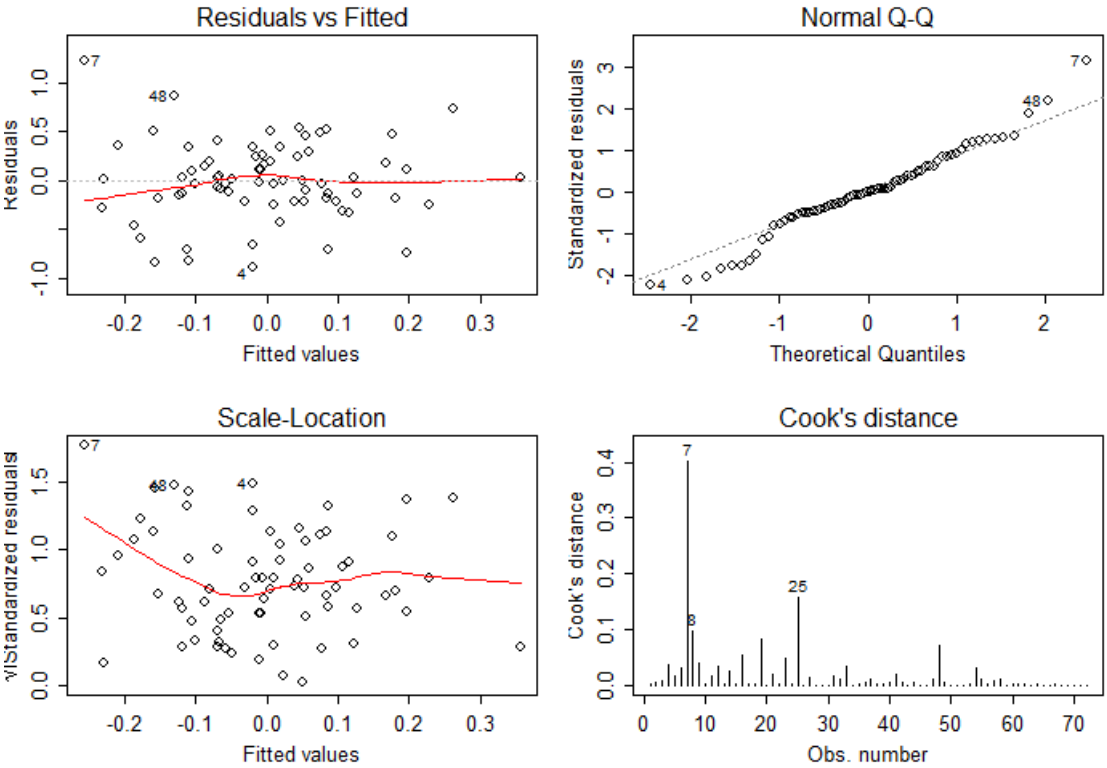
No. trials	Trial Size	R_h^2	IQR R_h^2	% Cover CIs	Low CI	Upp CI
5	10	0.879	0.066	100%	0.495	0.982
5	20	0.903	0.044	100%	0.673	0.974
5	40	0.902	0.039	99%	0.751	0.959
5	60	0.888	0.044	100%	0.768	0.946
5	100	0.902	0.038	96%	0.820	0.944
5	150	0.895	0.037	92%	0.830	0.933
5	200	0.901	0.033	94%	0.844	0.932
5	300	0.903	0.035	91%	0.862	0.930
10	10	0.864	0.062	100%	0.481	0.977
10	20	0.883	0.041	100%	0.642	0.967
10	40	0.892	0.032	100%	0.743	0.956
10	60	0.887	0.036	99%	0.771	0.946
10	100	0.890	0.035	96%	0.806	0.937
10	150	0.888	0.033	94%	0.824	0.930
10	200	0.892	0.037	88%	0.839	0.928
10	300	0.892	0.033	85%	0.851	0.923
20	10	0.861	0.033	100%	0.472	0.976
20	20	0.882	0.028	100%	0.640	0.966
20	40	0.883	0.025	100%	0.734	0.952
20	60	0.879	0.027	100%	0.761	0.942
20	100	0.885	0.025	100%	0.802	0.935
20	150	0.879	0.027	98%	0.813	0.924
20	200	0.879	0.034	94%	0.824	0.920
20	300	0.890	0.028	92%	0.847	0.921
30	10	0.856	0.029	100%	0.467	0.975
30	20	0.881	0.024	100%	0.642	0.965
30	40	0.881	0.021	100%	0.729	0.951
30	60	0.878	0.021	100%	0.761	0.942
30	100	0.884	0.024	100%	0.801	0.934
30	150	0.879	0.022	99%	0.814	0.925
30	200	0.872	0.028	98%	0.816	0.915
30	300	0.888	0.029	93%	0.846	0.920

Table C.11: *Ceiling: Simulation ordinal-ordinal: Results R_h^2. Where R_h^2=1 and R_{ht}^2=0.90 and the odds are proportional*

Appendix D: Case study

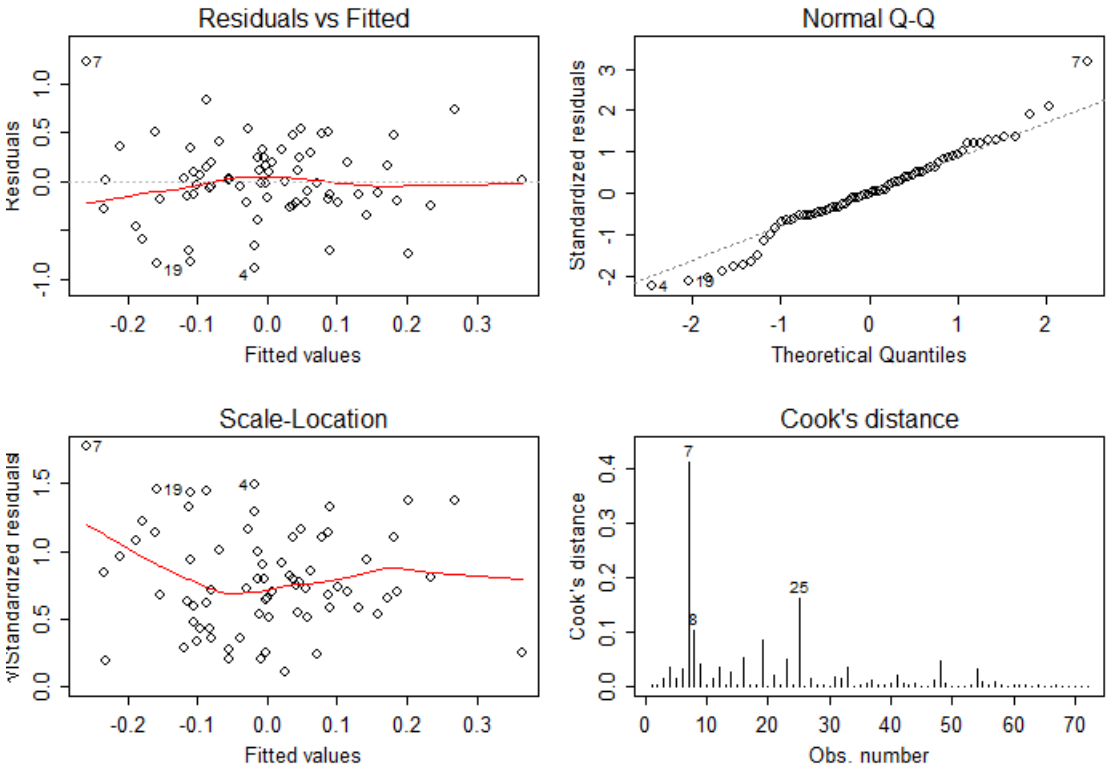
This appendix gives model diagnostic plots for second stage models of trial level surrogacy produced in the case study. In each case the diagnostic plots given are a: residual vs fitted values plot; QQ plot; standardized residual vs fitted values plot; and a Cook's distance plot. These diagnostic plots are given for second stage models for: the clinical results of CLOTs3; and the methodological setting where separation was ignored. The diagnostics plots are given for each surrogate in each setting

D.1. Diagnostics for clinical surrogacy assessment



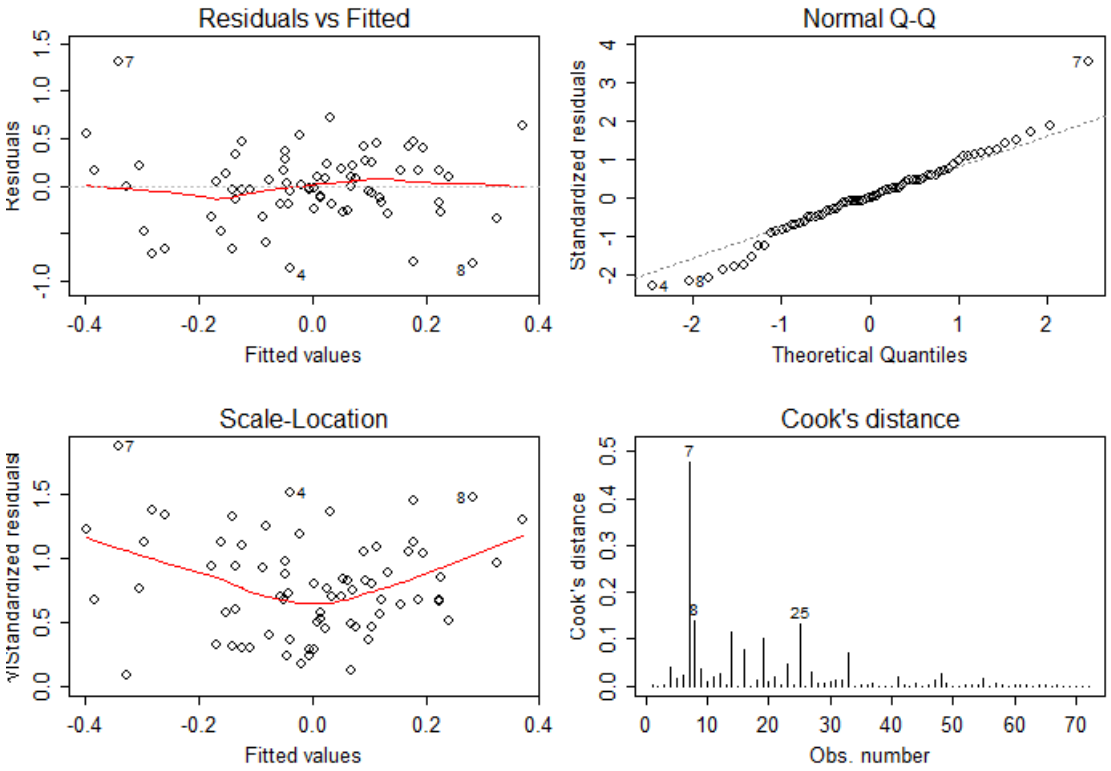
D. 1: *Binary-ordinal setting: Diagnostic plots for the second stage models of the surrogate evaluation of DVT by 30 days for OHS at six months.*

There was an indication that trial seven was an outlier but there appeared to be no particular reason for this and removal of the trial did not change the results. Hence this trial was retained in the analysis.



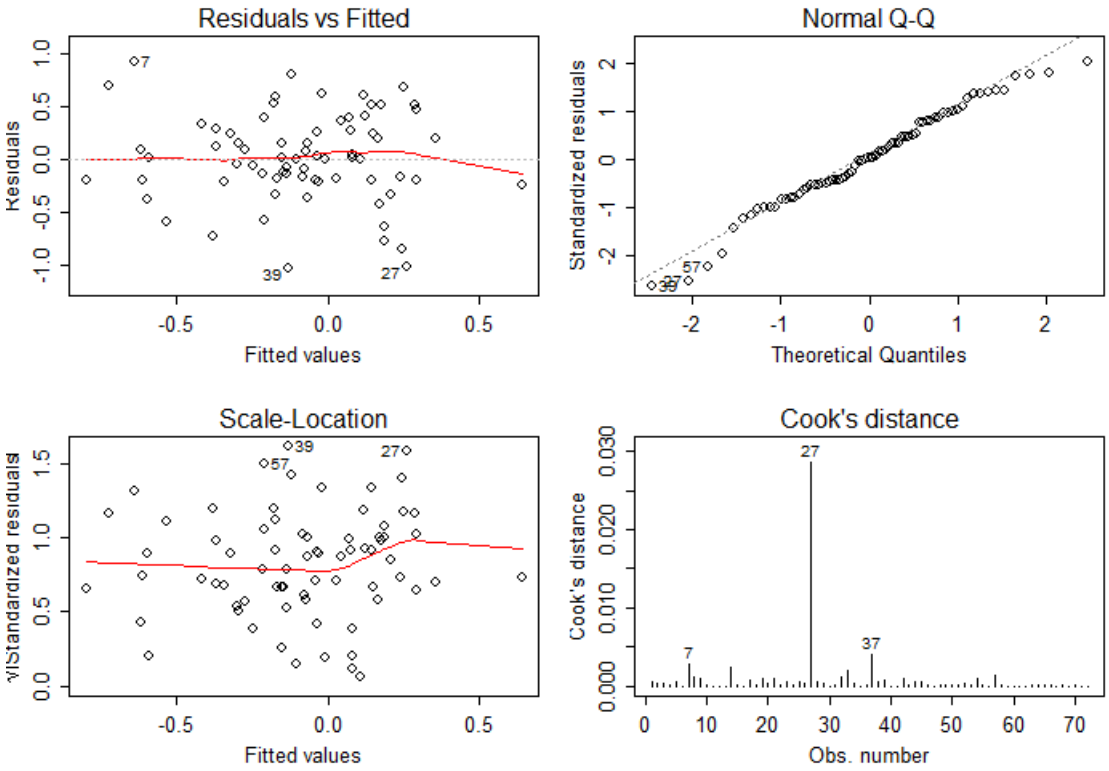
D. 2: *Binary-ordinal setting: Diagnostic plots for the second stage models of the surrogate evaluation of DVT and PE by 30 days for OHS at six months.*

There was an indication that trial seven was an outlier but there appeared to be no particular reason for this and removal of the trial did not change the results. Hence this trial was retained in the analysis.

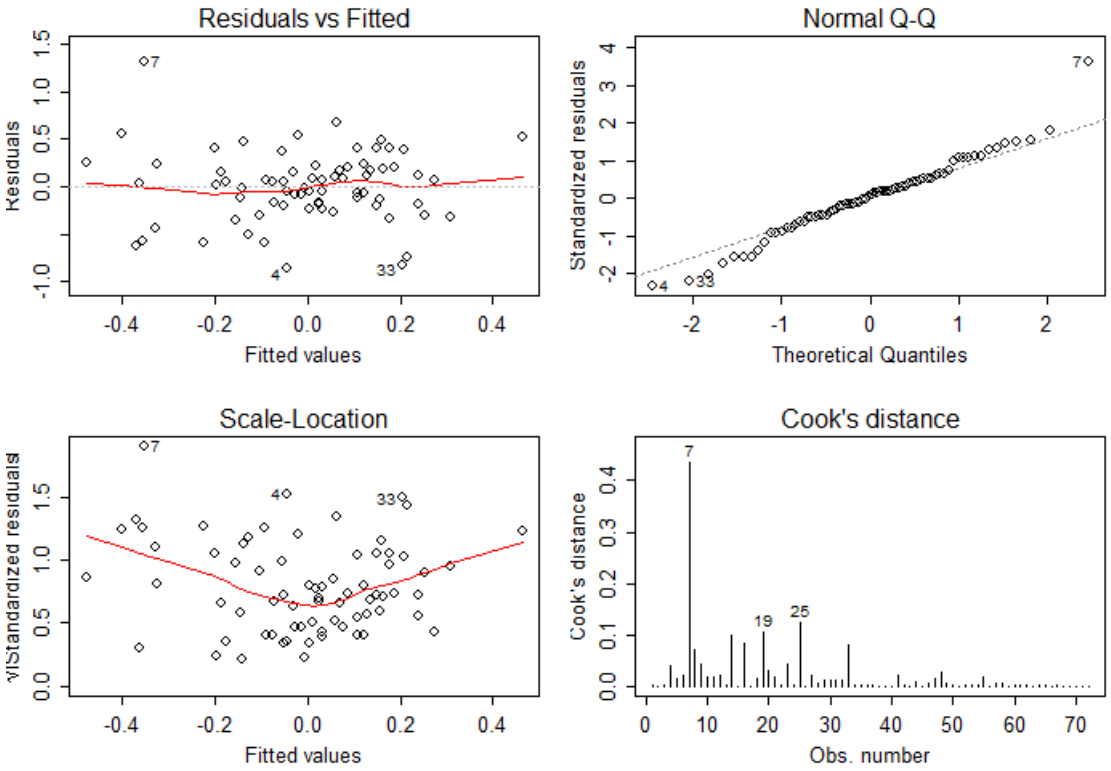


D. 3: *Binary-ordinal setting: Diagnostic plots for the second stage models of the surrogate evaluation of DVT, PE and Death by 30 days for OHS at six months.*

There was an indication that trial seven was an outlier but there appeared to be no particular reason for this and removal of the trial did not change the results. Hence this trial was retained in the analysis.



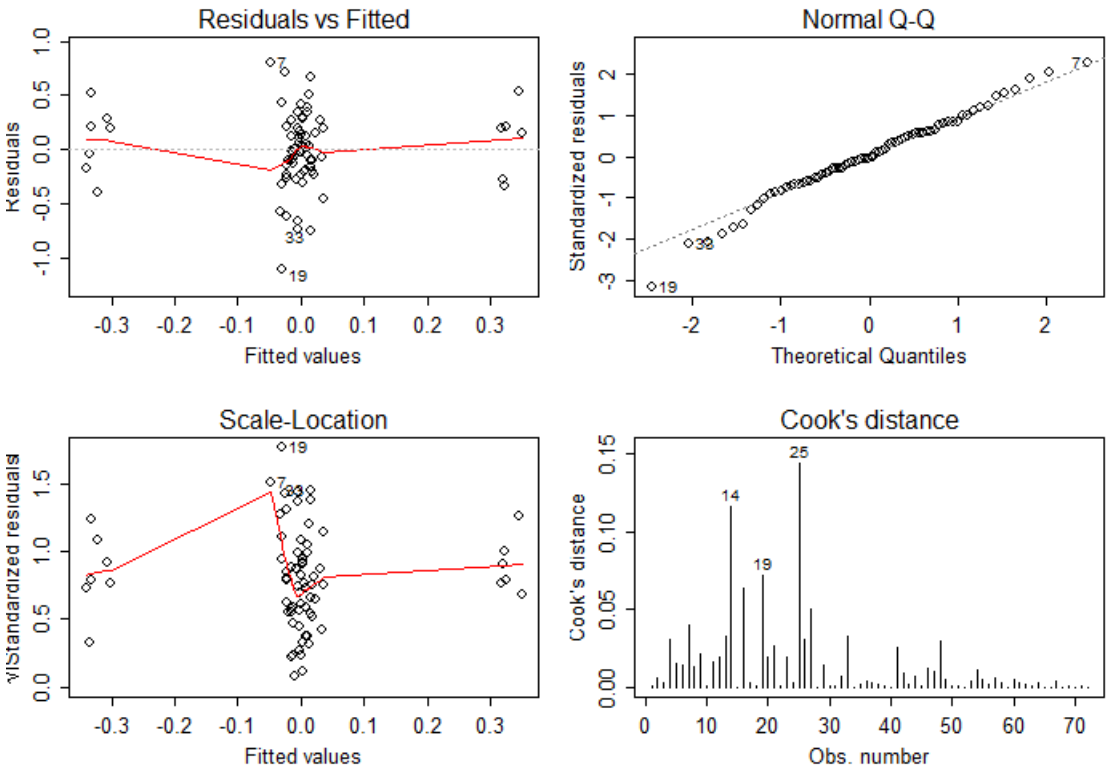
D. 4: *Ordinal-binary setting: Diagnostic plots for the second stage models of the surrogate evaluation oDVT by 30 days for binary survival at six months.*



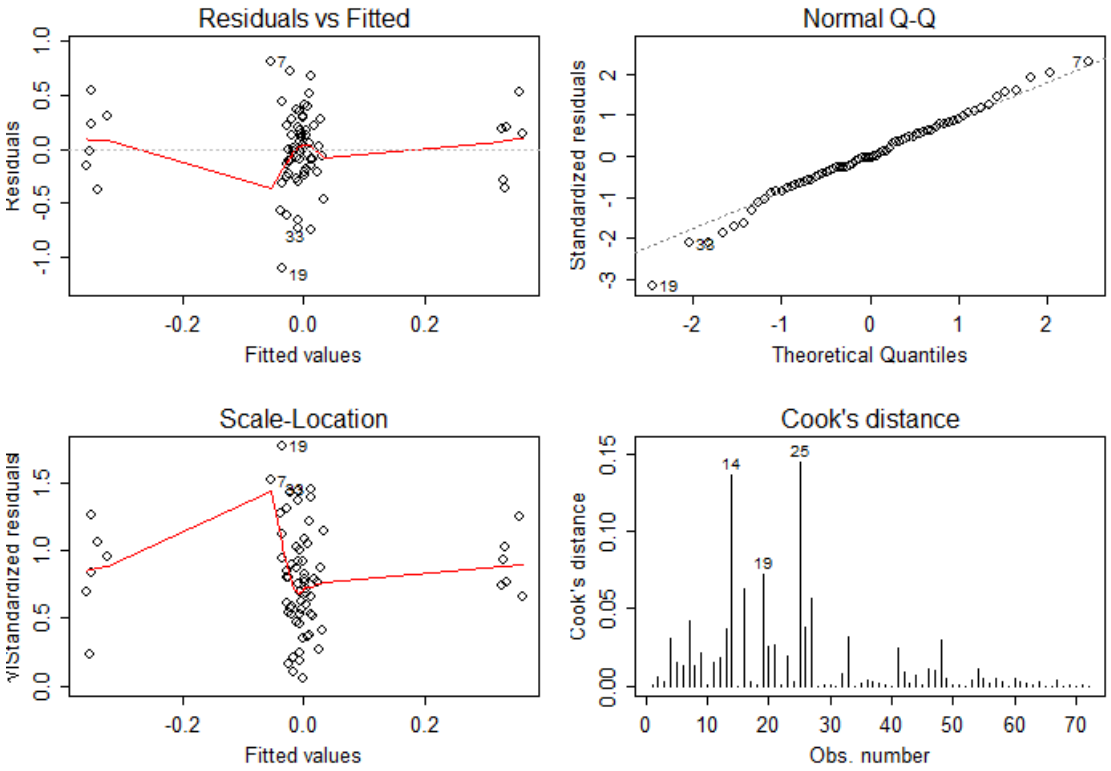
D. 5: *Ordinal-ordinal setting: Diagnostic plots for the second stage models of the surrogate evaluation oDVT by 30 days for OHS at six months.*

There was an indication that trial seven was an outlier but there appeared to be no particular reason for this and removal of the trial did not change the results. Hence this trial was retained in the analysis.

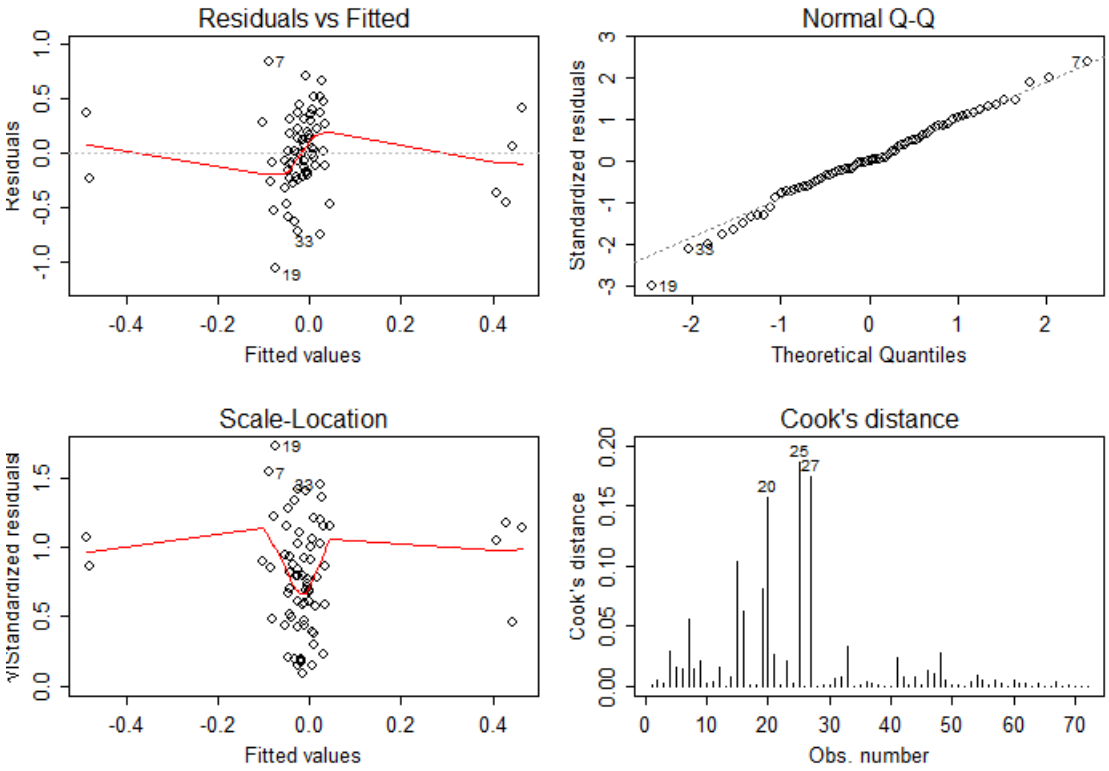
D.2. Diagnostics for regression where separation ignored



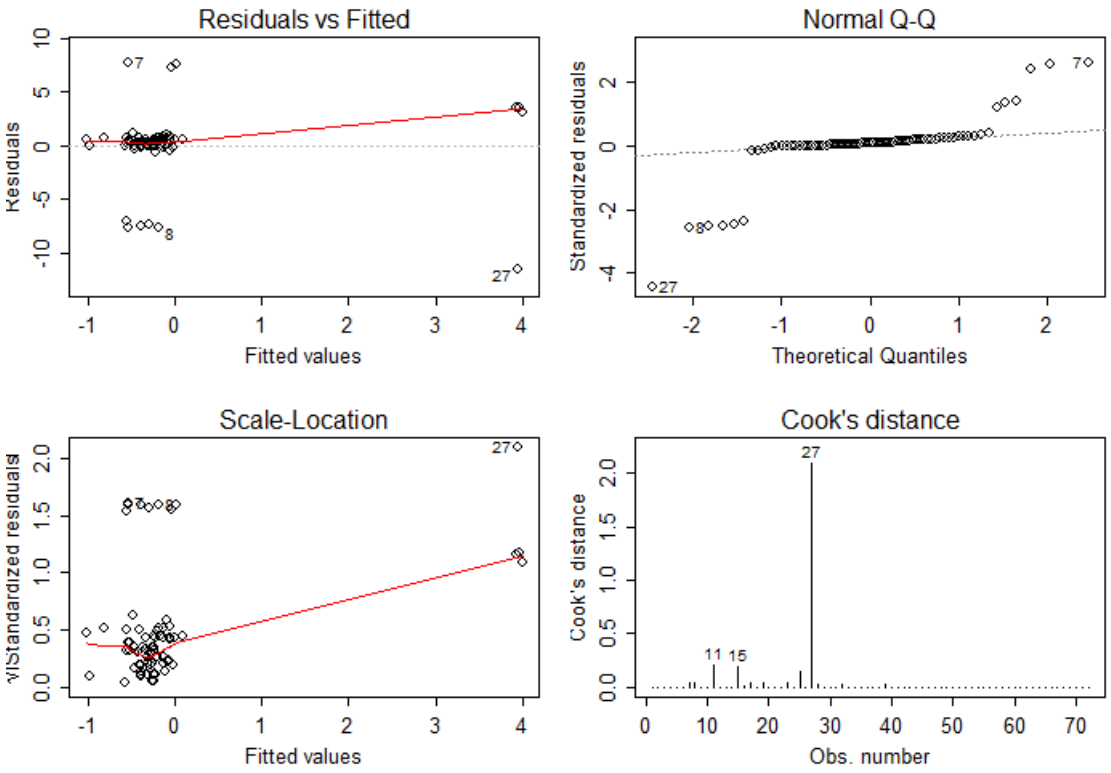
D. 6: *Binary-ordinal setting: Diagnostic plots for the second stage models of the surrogate evaluation of DVT by 30 days for OHS at six months where separation was ignored.*



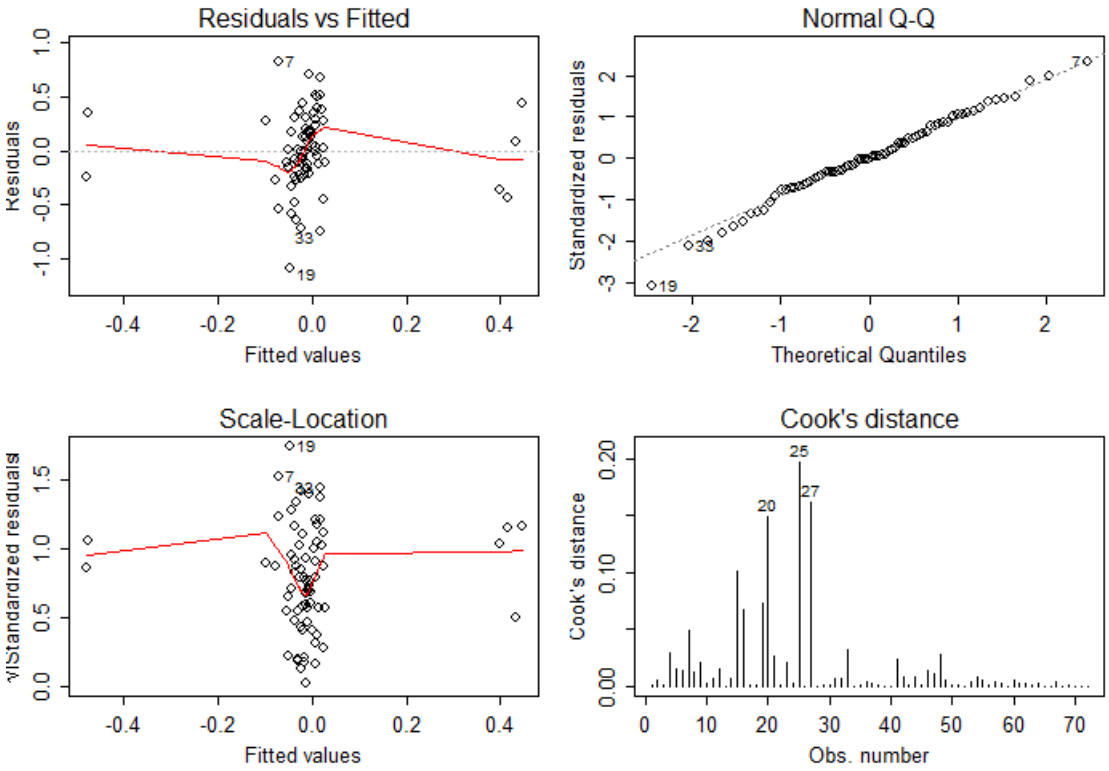
D. 7: *Binary-ordinal setting: Diagnostic plots for the second stage models of the surrogate evaluation of DVT and PE by 30 days for OHS at six months where separation was ignored.*



D. 8: *Binary-ordinal setting: Diagnostic plots for the second stage models of the surrogate evaluation of DVT, PE and Death by 30 days for OHS at six months where separation was ignored.*



D. 9: *Ordinal-binary setting: Diagnostic plots for the second stage models of the surrogate evaluation of oDVT by 30 days for survival at six months where separation was ignored.*



D. 10: Ordinal-ordinal setting: Diagnostic plots for the second stage models of the surrogate evaluation of oDVT by 30 days for OHS at six months where separation was ignored.