Separation and the information theory surrogate evaluation approach

Citation for published version:
Ensor, H & Weir, CJ 2021, 'Separation and the information theory surrogate evaluation approach: a penalized likelihood solution', *Pharmaceutical Statistics*. https://doi.org/10.1002/pst.2152

Digital Object Identifier (DOI):
10.1002/pst.2152

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Pharmaceutical Statistics

Received: 13 October 2020 | Revised: 10 May 2021 | Accepted: 28 June 2021

DOI: 10.1002/pst.2152

MAIN PAPER

WILEY

Separation and the information theory surrogate evaluation approach: A penalised likelihood solution

Hannah Ensor | Christopher J. Weir

Edinburgh Clinical Trials Unit, Usher Institute, University of Edinburgh, Edinburgh, UK

Correspondence
Hannah Ensor, Edinburgh Clinical Trials Unit, University of Edinburgh, Usher Institute, Nine Bioquarter, 9 Little France Road, Edinburgh EH16 4UX, UK.
Email: hannah.ensor@ed.ac.uk

Funding information
NHS Lothian via the Edinburgh Clinical Trials Unit; Medical Research Council Trials Methodology Research Studentship

Abstract

Surrogate evaluation is an important topic in clinical trials research, the use of a surrogate in place of a primary endpoint of interest is a common occurrence but also a contentious issue that is much debated. Statistical techniques to assess potential surrogates are closely scrutinised by the research community given the complexities of such an assessment. One such technique is the information theory surrogate evaluation approach which is well-established, practical and theoretically sound. In the context of discrete outcomes, we investigated issues of bias due to inefficiency, overfitting and separation (sparse data) that have not been recognised or addressed previously. The most serious cause of bias is separation in trial information. We outline the concerns surrounding this bias and conduct a simulation study to investigate whether a penalised likelihood technique provides an appropriate solution. We found that removing trials with separation from surrogacy evaluation resulted in a large amount of discarded data. Conversely, the penalised likelihood technique allows retention of all trial information and enables precise and reliable surrogate estimation. The information theory approach is a critical tool for conducting surrogate evaluation. This work strengthens the practical application of the information theory approach, allowing analyses to be adapted or the results summarised with appropriate caution to mitigate the biases highlighted. This is especially true where separation occurs. The adoption of the penalised likelihood technique into information theory surrogate evaluation is a useful addition that solves an issue likely to arise frequently in the context of categorical endpoints.

KEYWORDS

information theory, penalised likelihood, separation, surrogate endpoint, surrogate evaluation

1 | INTRODUCTION

Surrogates are measures of treatment effect that can be evaluated early and inform on the treatment effect on the primary outcome of interest. The use of a valid surrogate in place of the primary outcome offers potentially huge cost and time benefits. However, the use of an invalid surrogate could be extremely detrimental to the drug development process and to patient safety. Evaluating surrogates is as crucial as it is difficult: complexities in treatment mechanisms of action can mask potential inadequacies in a surrogate.

One well-established practical approach to surrogate evaluation is a multi-trial approach using information theory.[1,2] This approach generates estimates of surrogacy at two levels. Trial-level surrogacy quantifies the association between treatment effect estimates on the surrogate and true outcome in a trial, while individual-level surrogacy measures the correlation at the individual patient level after adjusting for treatment. The information theory approach has been extended to the case of continuous,[3] binary,[4] ordinal,[5] time-to-event[6] and longitudinal outcomes.[7] All of these settings have been thoroughly investigated via case studies and simulations. These multi-trial methodologies have also been used frequently in real applications to inform clinical trial practice[8] and there are calls for their use to be a requirement of regulatory bodies for studies investigating new drugs based on a surrogate.[9–11] Finally, SAS code and an R package help the applied researcher implement this methodology.[12]

We previously extended the information theory approach to the case of a binary surrogate and ordinal true outcome (the binary-ordinal setting).[5] We identified three forms of potential bias in assessing trial-level surrogacy. These were due to inefficiency, overfitting and the impact of separation in discrete outcomes. Underestimation (of strength of surrogacy) occurred when there were data available from a large number of trials, due to the loss of efficiency inherent in discrete outcomes and the two-stage nature of the modelling. Overestimation was present when only small numbers of trials were available, due to overfitting in the first stage of modelling. The most serious of the three issues identified was the impact of separation (e.g., a zero cell in a cross-tabulation of two binary outcomes). Discrete outcomes are very common in medical practice but the usual logistic regression analysis of these is biased in the presence of separation. These issues require thorough elucidation so that the analysis of discrete outcomes, and in particular the information theory approach to surrogate evaluation, can be optimised. We investigate the causes of the biases due to inefficiency, overfitting and separation, the form they take and the conditions and settings under which they are strongest and most prevalent. Finally, we offer a solution to the most serious form of bias identified, that which occurs in the presence of separation.

In Section 2 we summarise the information theory surrogate evaluation approach and show how it is applied in the binary-ordinal setting to evaluate trial-level surrogacy. In Section 3 we outline how bias has a serious effect on surrogate evaluation in the presence of separation and present a penalised likelihood technique as a solution. In Section 4 we discuss how inefficiency and overfitting affect estimation. In Section 5 we present a simulation study to explore issues of bias in more detail, with conclusions in Section 6.

2 | THE INFORMATION THEORY APPROACH

Since the biases identified only impact on trial-level surrogacy, this is the only information theory surrogate evaluation measure we derive in this section.

In what follows: Y represents a discrete random variable with values $k_b, b \in (1, \ldots, m_y)$ and probabilities of occurrence of each value p_b. We represent a putative surrogate as S, the treatment group indicator as Z and the true ordinal outcome as T. The categories of T are denoted by $w = 1, \ldots,$ W. In the multi trial context there are $i = 1, 2, \ldots, N$ trials, and $j = 1, 2, \ldots, n_i$ patients per trial.

Surrogate evaluation was previously proposed using a meta-analytical approach with a joint mixed model of the true and surrogate outcomes regressed on treatment. However, the model was found to be computationally burdensome.[13] Tibaldi et al.[14] suggested that a two stage fixed effects approach[13] would be preferable to the full mixed effects model as it is more computationally feasible and has only a minor loss of statistical efficiency (for normally distributed outcomes). This two-stage approach was found to work well in various settings (e.g., with binary, continuous or time-to-event outcomes). However, at the individual level different measures of association in different settings meant there was no consistent interpretation. The information theory approach[1] was developed to resolve this inconsistency.

Information theory[15] concerns information, choice and uncertainty in a draw from a random process. Entropy is a key concept that quantifies the amount of information gained from such a draw. Entropy can be expressed mathematically as $H(Y) = -\sum_{b=1}^{m_y} p_b \log(p_b)$, 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. A full list of the properties of entropy can be found in Shannon and Weaver.[15] An extension of the concept of entropy for continuous outcomes is a measure called entropy power (EP) which is used to compare random variables.[15]

Mutual information is a key concept that quantifies the amount of uncertainty in a variable expected to be reduced if information about another variable is known. It is defined as $I(X,Y)=H(Y)-H(Y|X)$, where $H(Y|X)$ is the conditional entropy of Y|X. In the case of surrogate outcomes, this quantity can be considered as the information in T that is shared by S.

Alonso and Molenberghs[1] proposed a trial level measure of surrogacy based on these concepts, called R^2_{ht}:

$$R^2_{\text{ht}}=1-\frac{\text{EP}(\beta|\alpha)}{\text{EP}(\beta)} \tag{1}$$

where α_i and β_i are the treatment effects in trial i on the surrogate and true outcome respectively. EP$(\beta|\alpha)$ is the entropy power of the distribution of β_i given the distribution ofα_i and EP(β) is the entropy power of the distribution of β_i. R^2_{ht} can be interpreted as the proportion of uncertainty in the treatment effects on T removed by adjusting for treatment effects on S. In the bivariate continuous setting, R^2_{ht} can be shown to reduce to the trial level surrogacy measure proposed under the meta-analytical approach. These concepts are consistent with the aims of surrogate evaluation since it is concerned with increasing our knowledge about the treatment effect on the true outcome using the surrogate.

In order to estimate R^2_{ht} Alonso and Molenberghs[1] suggested using the likelihood reduction factor, LRF, introduced by Alonso et al.[16] This provides consistent estimation of surrogacy; ranges in the unit interval; has a common interpretation across settings; and avoids evaluating high dimensional integrals and joint models for X and Y which would otherwise be required when fitting the models required by the information theoretic approach.

2.1 | Likelihood reduction factor: binary-ordinal setting

Here we estimate the likelihood reduction factor using the two stage approach outlined by Alonso and Molenberghs[1] but applied to the binary-ordinal setting described in Ensor and Weir.[5]

At the trial level, we focus on the treatment effects on the surrogate in relation to the treatment effects on the true outcome. At the first stage interest is in the intercept and treatment effects for each trial on the binary surrogate and ordinal true outcome, μ_{S_i}, μ_{T_i}, α_i and β_i respectively. These are found by regressing the surrogate and true outcome on treatment using the logistic regression and proportional odds models (2) and (3) respectively.

$$\text{logit}\left[P\left(S_{ij}=1\right)\right]=\mu_{S_i}+\alpha_i Z_{ij} \tag{2}$$

$$\text{logit}\left[P\left(T_{ij}\le w\right)\right]=\mu_{T_{w_i}}+\beta_i Z_{ij} \tag{3}$$

At the second stage the parameter effect estimates from stage one are used in two further models. These are: the intercept-only model (4) of treatment effects on the true outcome for each trial; and the treatment effects on the true outcome regressed on the intercept and treatment effect of the surrogate for each trial, see model (5).

$$\widehat{\beta}_i=\gamma_3+\varepsilon_i \tag{4}$$

$$\widehat{\beta}_i=\gamma_0+\gamma_1\widehat{\mu}_{S_i}+\gamma_2\widehat{\alpha}_{\mathbf{i}}+\varepsilon_i, \tag{5}$$

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 effects. Now we calculate the difference in -2*log-likelihood, denoted as G^2, between these two models to determine the LRF.

$$\text{LRF}=\widehat{R^2_{ht}}=1-\exp\left(-\frac{G^2}{N}\right) \tag{6}$$

Here N is the total number of trials. In this case, the difference in the -2*log-likelihood summarises the amount of information on the treatment effect estimates on the true outcome, $\widehat{\beta}_i$, explained by the addition to the model of treatment effect estimates on the surrogate, $\widehat{\boldsymbol{\alpha}}_{\boldsymbol{i}}$. Therefore, the LRF conceptually links to R^2_{ht} which is defined in Equation (1) as the 'proportion of uncertainty in the treatment effects on T removed by adjusting for treatment effects on S'.

Confidence intervals can be calculated using an approach by Kent,[17] the construction of these is detailed in Ensor and Weir.[5]

3 | SEPARATION

Where separation or quasi-complete separation of categorical variables occurs, there is no unique maximum likelihood.[18] Let us consider the case of two binary variables where one is regressed on the other as in Equation (2). Complete and quasi-complete separation relate to the existence of empty cells in the cross-tabulation of S and Z. In Table 1A–C respectively we present the case of: no separation—with no empty cells; complete separation—where the binary variable Z perfectly predicts S; and quasi-complete separation—where one cell is empty. The maximum likelihood estimate for two binary variables is:

$$\widehat{\varphi} = \log\left(\frac{A*D}{B*C}\right) \tag{7}$$

If a zero occurs in the denominator or numerator of 7 the function is undefined (i.e., $\widehat{\varphi} = \infty$ in the case of the denominator and $\widehat{\varphi} = \log(0)$ for the numerator). Therefore, there is no maximum likelihood in the presence of separation.[18] Quasi-complete separation can occur in the case of an ordinal variable regressed on a binary one, in a similar manner and with similar consequences as in the binary case. In the case of the information theory approach for the binary surrogate in model (2) $\widehat{\varphi} = \widehat{\alpha}_i$.

3.1 | Impact of separation on information theory surrogate evaluation

When complete or quasi-complete separation occurs this typically causes problems with maximum likelihood estimation for generalised linear models. In the typical scenario, the model iterates several times trying to converge.[18] The affected parameter estimate increases on each iteration, continuing to do so until a fixed iteration limit. Generally by this point, the parameter estimate will be large and its standard error very large.

TABLE 1 Examples of complete and quasi-complete separation in the binary and ordinal setting

A: No separation		
	Treatment	**Placebo**
Surrogate Y	A ≠ 0	B ≠ 0
Surrogate N	C ≠ 0	D ≠ 0
B: Complete separation		
	Treatment	**Placebo**
Surrogate Y	A ≠ 0	0
Surrogate N	0	D ≠ 0
C: Quasi-complete separation		
	Treatment	**Placebo**
Surrogate Y	A ≠ 0	B ≠ 0
Surrogate N	C ≠ 0	0

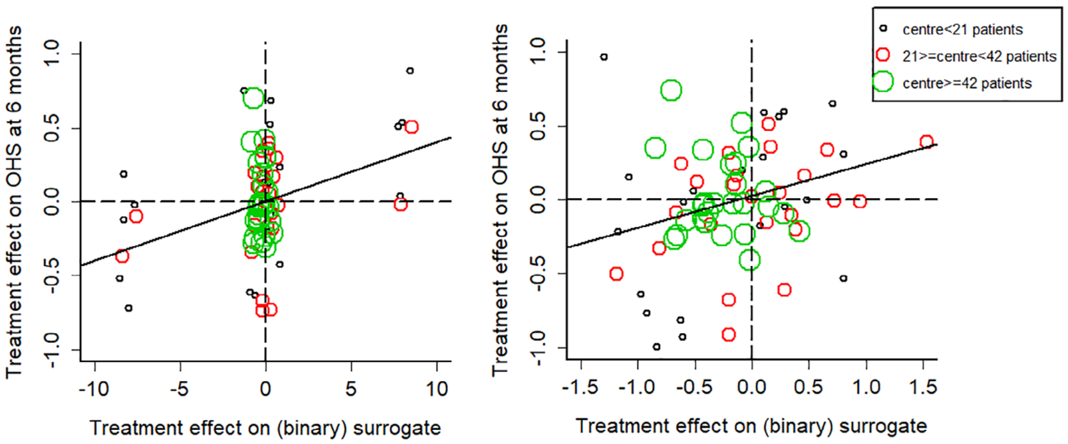

FIGURE 1 Illustration of the impact of separation based on the CLOTS3 trial[20] with a binary surrogate based on deep vein thrombosis and ordinal true outcome of the Oxford Handicap Scale. The left hand plot demonstrates the regression in the presence of separation ($R^2{}_{\mathrm{ht}} = 0.145$ 95% CI (0.027, 0.325) and the right gives the case where the penalised likelihood approach of Firth[19] is used ($R^2{}_{\mathrm{ht}} = 0.077$ 95% CI (0.003, 0.231)

At the first stage of surrogacy estimation, S and Z (both binary) and T and Z (one ordinal, one binary) are regressed on one another for each trial in models (2) and (3). This returns treatment effect estimates on the binary surrogate and ordinal true outcome. However, in the presence of separation these will be biased and since these estimates are used in modelling at stage two they tend to cause outlying points in the stage two regression. The LRF, Equation (6), is then based on models with potentially highly influential outliers. This leads to unreliable estimation of R^2_{ht}, with a tendency to underestimate the true value.

For a visual representation of the impact of separation see the results of a surrogacy assessment in Figure 1. The randomised trial Clots in Legs Or sTockings after Stroke (CLOTs3)[20] aimed to determine whether compression aids reduced the occurrence of deep vein thrombosis in immobile patients who had suffered stroke. We assessed if binary measures of deep vein thrombosis taken within 30 days of a stroke could be used as a surrogate in place of an ordinal measure of death and disability at 6 months post stroke. If centres (which can be used in place of trials in surrogacy evaluation[21]) with separation are retained in the usual information theory surrogacy assessment various outlying points can be seen at the left and right ends of the x-axis in the second stage of modelling. These are due to separation and potentially could strongly influence the regression parameter estimates. This issue occurs in several centres in this real life example. In the presence of separation $R^2{}_{\mathrm{ht}} = 0.145$, 95% CI (0.027, 0.325), and using the penalised likelihood approach $R^2{}_{\mathrm{ht}} = 0.077$, 95% CI (0.003, 0.231).

3.2 | Penalised likelihood solution to separation issues

Various possible solutions to the issue of separation[18] include deleting problematic variables; combining categories (in our case trials); reporting only the likelihood ratio statistics; using exact logistic regression, penalised maximum likelihood or Bayesian estimation.

Given the variables of interest, a desire to retain trial-specific information and the parameter estimation required in surrogacy evaluation only the latter three options are available to us. Allison[18] found that a Bayesian approach with uninformative priors led to convergence problems. Furthermore, parameter estimation is purportedly better for the penalised maximum likelihood than exact logistic regression.[22]

The penalised likelihood technique of Firth[19] was originally introduced to reduce bias in maximum likelihood estimates in logistic regression. In particular, it applies to small samples where bias increases away from zero and infinite parameter estimates in the case of separation can be thought of as an extreme example.[22]

If we have a scalar parameter θ of an exponential family model $l(\theta) = t\theta - K(\theta)$, the score function $U(\theta) = l'(\theta) = t - K'(\theta)$, that is, the sufficient statistic t affects the location but not the shape of $U(\theta)$. At the true value of θ we have $E(U(\theta)) = 0$ therefore the score function is unbiased. It is also true that if the score function is linear in θ then $E\left(\widehat{\theta}\right) = \theta$. However, when the score function is curved in θ, as is the case under separation, Firth states that the unbiasedness of the score function and this curvature imposes a bias in $\widehat{\theta}$ so that $E\left(\widehat{\theta}\right) \neq \theta$.

Firth[19] favoured a systematic modification to the score function (adding a bias term) to prevent bias in $\widehat{\theta}$ rather than correcting an already biased estimate of θ.

Specifically, Firth employed the following notation for the vector of parameters θ_r, where r = 1, ..., p, whose maximum likelihood estimates are usually determined from the score function equation $\frac{\delta \mathrm{Logl}}{\delta\theta} \equiv U(\theta_r) = 0$. For the exponential family, Firth suggested a modified score equation: $U^*(\theta_r) \equiv U(\theta_r) + \frac{\delta}{\delta\theta_r}\left\{\frac{1}{2}\log|i(\theta)|\right\}$, where $i(\theta)$ is the information matrix of θ. This equates to a penalised likelihood equation of $L^*(\theta) = L(\theta)|i(\theta)|^{1/2}$ where the bias term $|i(\theta)|^{1/2}$ is the Jeffreys invariant prior.[23] The influence of this bias term is asymptotically negligible. Firth[19] showed that this technique removed the overall bias in parameter estimation.

Heinze and Schemper[22] applied the technique of Firth[19] to deal with instances of separation in the logistic regression context. Under assessment Heinze and Schemper[22] showed that it was 'an ideal solution to separation' as it produces finite parameter estimates that are superior overall to those from alternative methods. Therefore, we apply the penalised likelihood technique of Firth[19] to resolve surrogacy estimation issues which result from separation.

Penalised likelihood techniques may be implemented for generalised linear models using the `logistf` command in the `logistf` package and `pordlogist` in package `OrdinalLogisticBioplot`[24] in R.

4 | OVERFITTING AND INEFFICIENCY IN THE INFORMATION THEORY APPROACH

Besides separation, we also identified issues of underestimation and overestimation previously.[5] Underestimation worsened as the number of trials increased, presumably due to inefficiency of the two-stage approach compounded by the presence of discrete outcomes. We found that the general literature supported this assessment: Molenberghs et al.[25] investigated the partitioning of a large dataset by applying a logistic regression to each partition to gain multiple estimates of the parameter of interest. The mean of these parameters was then compared to the estimate gained from the model using all the data. They showed that inefficiency occurred when the number of partitions (in our case trials) was large compared to the size of the partitions. On top of this, parameter estimation is less efficient if binary or ordinal outcomes are used in place of continuous outcomes.[26,27]

Overestimation occurred in the presence of weak surrogacy and a small number of trials. This overestimation was due to overfitting of the regression at the second stage of modelling to a small number of data points (one for each trial). Since the second stage models affected by overfitting are normal linear models, irrespective of the type of outcome being studied, such overfitting would be expected to be present in all settings. The classical R^2 measure of the coefficient of determination is known to be biased and inflated particularly in the case of small sample sizes and/or too many predictors. The information theory approach was introduced to address issues with a unified interpretation at the individual level. At the trial level the calculation of R^2_{ht} has been shown to be consistent with the classical R^2 measure.[28] Therefore the R^2_{ht} also suffers this bias.

The classical R^2 can be adjusted through the calculation of the required shrinkage to provide an unbiased estimator of the population R^2, which we denote R^2_{adjC}. Unbiased estimation of the population surrogacy strength is the primary focus of surrogacy evaluation. In order to assess overfitting of R^2_{ht} we will also present R^2_{adjC} in simulations.

Methodologists have previously discussed a full and reduced model at the second stage of analysis. The full model is shown in Equation (5), and has two explanatory variables. The reduced model is the same equation without the trial specific intercept estimates, $\widehat{\mu}_{S_i}$. The calculation of R^2_{ht} for the reduced model proceeds in the same manner as for the full model but based on the reduced regression. Tibaldi et al.[14]explored simplified means of surrogacy assessment. This paper concluded that in general the full model confers a small benefit and should be used in practice, and this has since been the convention. Previous simulations have only focused on strong surrogacy, predominantly $R^2_{\mathrm{ht}} = 0.9$, which explains why the issue of overestimation was not identified previously. A model based on fewer explanatory variables is likely to suffer less inflation of R^2_{ht}, therefore, we revisit this convention by presenting results of a reduced R^2, namely $R^2_{\mathrm{ht.R}}$. For completeness we will also include the adjusted estimate of the reduced model $R^2_{\mathrm{adjC.R}}$ in our simulations.

5 | SIMULATION

We investigated the practical worth of the penalised likelihood technique via a simulation study using R, based on the approach of Tilahun et al.[4] Various scenarios were simulated to study the R^2_{ht} surrogacy measure estimation. Trial sizes were set to 10, 20, 60, 100, and 300 patients. There were 5, 10, 20 or 30 trials in each simulated data set and 250 datasets simulated for each scenario. (Note that surrogacy data from multiple individual trials with the same design and treatment classes are not commonly available, therefore researchers tend to use centres within trial in place of trials, and evaluate surrogacy using data from multiple centres instead.[5] Centres within trials can have various different sizes from small to large; hence the above simulation trial size scenarios are representative of realistic settings that occur in practice). We present the mean point estimates and the variance. The joint mixed model in eight forms the basis for the data generation:

$$S_{ij} = \mu_S + m_{S_i} + \alpha Z_{ij} + a_i Z_{ij} + \varepsilon_{S_{ij}} \tag{9}$$

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

$$D = 3\begin{pmatrix} 1 & 075 & 0 & 0 \\ 0.75 & 1 & 0 & 0 \\ 0 & 0 & 1 & \rho \\ 0 & 0 & \rho & 1 \end{pmatrix}, \text{hence } R^2_{\text{ht}} = \rho^2, \sum = 3\begin{pmatrix} 1 & \psi \\ \psi & 1 \end{pmatrix}, \text{hence } R^2_h = \psi^2.$$

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 ith trial respectively. Error terms are jointly distributed, $\left(\varepsilon_{S_{ij}}, \varepsilon_{T_{ij}}\right) \sim \mathrm{N}(0, \sum)$ and random effects, $(m_{S_i}, m_{T_i}, a_i, b_i)^{\mathrm{T}} \sim \mathrm{N}(0, \mathrm{D})$. Intercept and treatment effect parameters for S and T were set to $\mu_s = 0.50$, $\mu_T = 0.45$, $\alpha = 0.05$, and $\beta = 0.03$. Surrogacy was simulated to be: strong at both trial level ($R^2_{\mathrm{ht}} = \rho^2 = 0.90$) and individual level surrogacy ($R^2_{\mathrm{h}} = \psi^2 = 0.64$); or weak on both measures, with $R^2_{\mathrm{ht}} = \rho^2 = 0.30$ and $R^2_{\mathrm{h}} = \psi^2 = 0.30$. After simulating continuous S and T these were then dichotomised or categorised to represent a binary S and a seven category ordinal T. In order to create a binary surrogate outcome, a simulated continuous surrogate was dichotomised at the mean. This was in keeping with previous publications in this field.[4,29,30] To create ordinal outcomes the continuous variables were categorised using six evenly spaced cut off points, determined according to the quantiles of the true outcome variable.

Two techniques to deal with separation were investigated. The first was to apply a penalised likelihood technique that allowed trials in which separation occurred to be retained in analysis. The second was to remove trials where separation occurred. If fewer than three trials remained following trial removal the simulation was set to return a null value. For the removal technique the simulation was run until 250 datasets were simulated with three or more trials available.

In order to investigate the other underestimation and overestimation biases discussed in Section 4, results for the binary-ordinal setting are compared to the information rich continuous-continuous setting. To further investigate whether the underestimation seen for large trial sizes is due to inefficiency a much larger trial size of 3000 patients per trial was also simulated.

5.1 | Results

For both strong and weak surrogacy settings, the penalised likelihood technique was compared to the removal of trials technique. The penalised likelihood technique was based on the full dataset in each case, whereas the trial removal technique was often based on a much reduced dataset due to the removal of data from trials in which separation occurred. For small numbers of patients per trial, often fewer than half of the trials were retained in analysis; see Tables 2 and 3. Furthermore, sometimes where there were fewer than three trials available for analysis under the trial removal technique the second stage of modelling could not proceed: in some scenarios this occurred up to 90% of the time. Even where there were large numbers of patients per trial (e.g., 300 patients) the median number of trials retained in the removal technique was generally lower than the number simulated, see Tables 2 and 3. This represents a frequent loss of information when the trial removal technique is implemented and shows how often separation can occur. The penalised likelihood technique by comparison allows the retention of all trials in the analysis.

Comparing the R^2_{ht} between the penalised likelihood and the trial removal techniques is not straightforward given the conflicting issues of bias and the inclusion of different numbers of trials under each method. For instance, in Table 2 where $\mathrm{R}^2_{\mathrm{ht}}$ is set to 0.90, the number of trials to 30 and the size of the trials to 10, the mean R^2_{ht} values are comparable which seems unlikely. The penalised likelihood and trial removal technique mean estimates are respectively $R^2_{\mathrm{ht}} = 0.517$ and $R^2_{\mathrm{ht}} = 0.514$. The median number of trials contributing to the trial removal technique in this setting is

TABLE 2 Simulation study: Mean $R^2{}_{\mathrm{ht}}$ estimates based on 250 simulations for each scenario, where true values set to: $R^2{}_{\mathrm{ht}} = 0.90$ and $R^2{}_{\mathrm{h}} = 0.64$

		Penalised likelihood technique				Trial removal technique					
No. trials	Trial size	Mean R^2_{ht}	Var R^2_{ht}	Mean R^2_{adjC}	Var R^2_{adjC}	Mean R^2_{ht}	Var R^2_{ht}	Mean R^2_{adjC}	Var R^2_{adjC}	% Failures	Median No. trialsavailable
5	10	0.7266	0.0545	0.4516	0.2197	0.7174	0.0820	0.2594	0.5634	90%	3
5	20	0.8047	0.0391	0.6094	0.1565	0.8064	0.0475	0.4871	0.3721	59%	4
5	60	0.8660	0.0251	0.7319	0.1003	0.8705	0.0290	0.6957	0.1717	13%	5
5	100	0.8799	0.0233	0.7598	0.0930	0.8803	0.0249	0.7317	0.1335	6%	5
5	300	0.9053	0.0172	0.8103	0.0686	0.9080	0.0148	0.8124	0.0593	2%	5
10	10	0.5810	0.0409	0.4606	0.0681	0.7043	0.0702	0.3358	0.3764	45%	4
10	20	0.7120	0.0275	0.6298	0.0454	0.7217	0.0479	0.5240	0.1646	6%	6
10	60	0.8294	0.0119	0.7806	0.0197	0.8157	0.0206	0.7458	0.0411	0%	9
10	100	0.8362	0.0118	0.7894	0.0195	0.8304	0.0148	0.7734	0.0264	0%	9
10	300	0.8738	0.0093	0.8376	0.0154	0.8718	0.0093	0.8324	0.0159	0%	10
20	10	0.5561	0.0213	0.5033	0.0267	0.5982	0.0608	0.3656	0.1627	2%	7
20	20	0.6624	0.0155	0.6226	0.0194	0.6568	0.0259	0.5775	0.0398	0%	12
20	60	0.7850	0.0090	0.7597	0.0112	0.7920	0.0084	0.7617	0.0110	0%	17
20	100	0.8265	0.0059	0.8061	0.0074	0.8307	0.0056	0.8077	0.0072	0%	18
20	300	0.8667	0.0033	0.8510	0.0041	0.8647	0.0036	0.8479	0.0046	0%	19
30	10	0.5167	0.0182	0.4806	0.0210	0.5143	0.0518	0.3673	0.0940	1%	10
30	20	0.6600	0.0107	0.6348	0.0124	0.6365	0.0226	0.5885	0.0292	1%	19
30	60	0.7870	0.0053	0.7713	0.0061	0.7847	0.0059	0.7651	0.0070	0%	25
30	100	0.8146	0.0046	0.8009	0.0054	0.8207	0.0043	0.8055	0.0051	0%	27
30	300	0.8578	0.0030	0.8472	0.0035	0.8582	0.0029	0.8471	0.0033	0%	29

Note: Comparing penalised likelihood technique against trial removal technique (trial removal technique results include the % of times the calculation of $R^2{}_{\mathrm{ht}}$ was not possible because fewer than three trials were retained and the median number of trials available for analysis when it was). R^2_{ht} is the information theory R^2 measure of the full model; $R^2_{\mathrm{adjC}}$is the adjusted coefficient of determination of the full model.

TABLE 3 Simulation study: Mean $R^2{}_{\text{ht}}$ estimates based on 250 simulations for each scenario, where true values set to: $R^2{}_{\text{ht}} = 0.30$ and $R^2{}_{\text{h}} = 0.30$

		Penalised likelihood technique				**Trial removal technique**					
No. trials	**Trial size**	Mean $\boldsymbol{R^2_{\text{ht}}}$	Var $\boldsymbol{R^2_{\text{ht}}}$	Mean $\boldsymbol{R^2_{\text{adjC}}}$	Var $\boldsymbol{R^2_{\text{adjC}}}$	Mean $\boldsymbol{R^2_{\text{ht}}}$	Var $\boldsymbol{R^2_{\text{ht}}}$	Mean $\boldsymbol{R^2_{\text{adjC}}}$	Var $\boldsymbol{R^2_{\text{adjC}}}$	**% failures**	**Median No. trials available**
5	10	0.5717	0.0823	0.1386	0.3359	0.6492	0.097	0.0453	0.7192	92%	3
5	20	0.5928	0.0768	0.1856	0.3073	0.6802	0.0762	0.1866	0.5004	59%	4
5	60	0.5686	0.0772	0.1372	0.3087	0.6115	0.0875	0.0832	0.5237	13%	4
5	100	0.6122	0.079	0.2244	0.3161	0.6326	0.0794	0.1839	0.4470	6%	5
5	300	0.600	0.0773	0.1987	0.3080	0.6206	0.078	0.2122	0.3508	2%	5
10	10	0.3597	0.0472	0.176	0.0781	0.6005	0.0848	0.066	0.5203	53%	4
10	20	0.3874	0.0478	0.2124	0.0790	0.503	0.0825	0.1621	0.2570	7%	6
10	60	0.3884	0.0438	0.2137	0.0725	0.423	0.0604	0.199	0.1253	0%	9
10	100	0.422	0.0471	0.2569	0.0778	0.4187	0.0521	0.2254	0.0938	0%	9
10	300	0.4087	0.0525	0.2375	0.0874	0.4111	0.0509	0.2311	0.0867	0%	10
20	10	0.2381	0.0197	0.1479	0.0248	0.3957	0.075	0.0255	0.2415	7%	7
20	20	0.2858	0.0261	0.2016	0.0326	0.3104	0.0387	0.1506	0.0583	0%	12
20	60	0.3413	0.0263	0.2638	0.0329	0.3332	0.0298	0.2355	0.0389	0%	17
20	100	0.3258	0.0268	0.2464	0.0335	0.308	0.0287	0.2151	0.0370	0%	18
20	300	0.3642	0.0263	0.2889	0.0329	0.3473	0.0288	0.2662	0.0365	0%	19
30	10	0.2116	0.0142	0.1529	0.0164	0.3195	0.0497	0.0893	0.0892	2%	10
30	20	0.2712	0.0158	0.2171	0.0183	0.2547	0.0243	0.1531	0.0321	1%	18
30	60	0.3173	0.0189	0.2667	0.0218	0.2996	0.0212	0.2355	0.0253	0%	25
30	100	0.3157	0.0173	0.265	0.0199	0.3041	0.0195	0.2455	0.0230	0%	27
30	300	0.3354	0.0174	0.2859	0.0201	0.3231	0.0168	0.2704	0.0196	0%	29

Note: Comparing penalised likelihood technique against trial removal technique (trial removal technique results include the % of times the calculation of $R^2{}_{\text{ht}}$ was not possible because fewer than three trials were retained and the median number of trials available for analysis when it was). R^2_{ht} is the information theory R^2 measure of the full model; R^2_{adjC} is the adjusted coefficient of determination of the full model.

TABLE 4 Simulation study: Mean R^2 estimates based on 250 simulations for each scenario, in binary-ordinal and continuous-continuous setting where true values set to: $R^2{}_{\text{h}} = 0.64$ and $R^2{}_{\text{ht}} = 0.90$

		Binary-ordinal								Continuous-continuous							
		R^2_{ht}		R^2_{adjC}		$R^2_{\text{ht.R}}$		$R^2_{\text{adjC.R}}$		R^2_{ht}		R^2_{adjC}		$R^2_{\text{ht.R}}$		$R^2_{\text{adjC.R}}$	
Number of trials	Trial size	Mean	Var.	Mean	Var.	Mean	Var.	Mean	Var.	Mean	Var.	Mean	Var.	Mean	Var.	Mean	Var.
5	10	0.7266	0.0545	0.4516	0.2197	0.5715	0.0776	0.4283	0.1380	0.855583	0.029847	0.710404	0.119601	0.797578	0.04205	0.729903	0.074744
5	20	0.8047	0.0391	0.6094	0.1565	0.6859	0.0649	0.5812	0.1155	0.88433	0.02153	0.768661	0.086119	0.828194	0.032369	0.770926	0.057545
5	60	0.8660	0.0251	0.7319	0.1003	0.7895	0.0364	0.7193	0.0648	0.907047	0.015407	0.814095	0.061627	0.862196	0.026045	0.816261	0.046302
5	100	0.8799	0.0233	0.7598	0.0930	0.8112	0.0375	0.7483	0.0667	0.915196	0.013675	0.830392	0.054701	0.87047	0.02537	0.827294	0.045102
5	300	0.9053	0.0172	0.8103	0.0686	0.8449	0.0299	0.7923	0.0543	0.910438	0.01049	0.820876	0.04196	0.860733	0.021945	0.814311	0.039013
5	3000	0.932339	0.007743	0.864448	0.030931	–	–	–	–	–	–	–	–	–	–	–	–
10	10	0.5810	0.0409	0.4606	0.0681	0.5141	0.0474	0.4530	0.0601	0.834907	0.011087	0.787654	0.018341	0.812585	0.013085	0.789121	0.016565
10	20	0.7120	0.0275	0.6298	0.0454	0.6609	0.0319	0.6186	0.0404	0.856672	0.010821	0.815721	0.017887	0.836655	0.012505	0.816236	0.015827
10	60	0.8294	0.0119	0.7806	0.0197	0.7936	0.0151	0.7678	0.0192	0.884107	0.008541	0.850995	0.014119	0.868266	0.010377	0.851799	0.013133
10	100	0.8362	0.0118	0.7894	0.0195	0.8024	0.0139	0.7777	0.0176	0.895513	0.005478	0.865659	0.009056	0.88032	0.006829	0.86536	0.008643
10	300	0.8738	0.0093	0.8376	0.0154	0.8526	0.0107	0.8341	0.0136	0.901975	0.005093	0.873968	0.008418	0.888473	0.006101	0.874532	0.007721
10	3000	0.896224	0.004887	0.866374	0.008089	–	–	–	–	–	–	–	–	–	–	–	–
20	10	0.5561	0.0213	0.5033	0.0267	0.5218	0.0219	0.4950	0.0245	0.822049	0.005868	0.801078	0.007334	0.810716	0.006508	0.800181	0.007254
20	20	0.6624	0.0155	0.6226	0.0194	0.6364	0.0162	0.6162	0.0180	0.859758	0.004771	0.843258	0.005959	0.850491	0.005194	0.842185	0.005787
20	60	0.7850	0.0090	0.7597	0.0112	0.7677	0.0102	0.7548	0.0113	0.882845	0.003162	0.869062	0.00395	0.876412	0.003357	0.869546	0.003741
20	100	0.8265	0.0059	0.8061	0.0074	0.8089	0.0073	0.7983	0.0081	0.887943	0.003322	0.87476	0.004149	0.88167	0.003573	0.875096	0.003982
20	300	0.8667	0.0033	0.8510	0.0041	0.8526	0.0040	0.8444	0.0045	0.899765	0.002461	0.887973	0.003075	0.894197	0.002487	0.88832	0.002771
20	3000	0.886237	0.002315	0.872743	0.002898	–	–	–	–	–	–	–	–	–	–	–	–
30	10	0.5167	0.0182	0.4806	0.0210	0.4938	0.0191	0.4756	0.0205	0.822125	0.00399	0.80892	0.004605	0.81554	0.004102	0.808938	0.004401
30	20	0.6600	0.0107	0.6348	0.0124	0.6422	0.0119	0.6295	0.0127	0.852071	0.003179	0.841113	0.003668	0.847055	0.003339	0.841593	0.003582
30	60	0.7870	0.0053	0.7713	0.0061	0.7756	0.0056	0.7676	0.0060	0.882195	0.002081	0.873469	0.0024	0.877888	0.002124	0.873527	0.002278
30	100	0.8146	0.0046	0.8009	0.0054	0.8036	0.0050	0.7966	0.0054	0.891815	0.001779	0.883802	0.002053	0.888272	0.001862	0.884282	0.001998
30	300	0.8578	0.0030	0.8472	0.0035	0.8490	0.0034	0.8436	0.0037	0.898883	0.001372	0.891393	0.001582	0.894876	0.001449	0.891121	0.001555
30	3000	0.886488	0.002068	0.878015	0.002388	–	–	–	–	–	–	–	–	–	–	–	–

Note: R^2_{ht} is the information theory R^2 measure of the full model; R^2_{adjC} is the adjusted coefficient of determination of the full model; $R^2_{\text{ht.R}}$ is the information theory R^2 measure of the reduced model; $R^2_{\text{adjC.R}}$ is the adjusted coefficient of determination of the reduced model.

Abbreviation: IQR, interquartile range.

TABLE 5 Mean R^2 estimates based on 250 simulations for each scenario, in continuous-continuous setting where true values set to: $R^2_{\text{h}} = 0.30$ and $R^2_{\text{ht}} = 0.30$

		Binary-ordinal								Continuous-continuous							
		R^2_{ht}		R^2_{adjC}		$R^2_{\text{ht.R}}$		$R^2_{\text{adjC.R}}$		R^2_{ht}		R^2_{adjC}		$R^2_{\text{ht.R}}$		$R^2_{\text{adjC.R}}$	
Number of trials	**Trial size**	**Mean**	**Var.**	**Mean**	**Var.**	**Mean**	**Var.**	**Mean**	**Var.**	**Mean**	**Var.**	**Mean**	**Var.**	**Mean**	**Var.**	**Mean**	**Var.**
5	10	0.5717	0.0823	0.1386	0.3359	0.3461	0.0773	0.1271	0.1378	0.614048	0.064411	0.228095	0.257644	0.414526	0.085484	0.219368	0.151971
5	20	0.5928	0.0768	0.1856	0.3073	0.3852	0.0882	0.1803	0.1568	0.597871	0.08141	0.195741	0.325642	0.440481	0.091865	0.253975	0.163316
5	60	0.5686	0.0772	0.1372	0.3087	0.3893	0.0863	0.1858	0.1534	0.625916	0.079343	0.251832	0.317371	0.412063	0.088765	0.216084	0.157805
5	100	0.6122	0.079	0.2244	0.3161	0.4037	0.0938	0.2049	0.1668	0.592625	0.081187	0.18525	0.324747	0.41059	0.085342	0.21412	0.151719
5	300	0.600	0.0773	0.1987	0.3080	0.4070	0.0892	0.2079	0.1588	0.622413	0.081851	0.244825	0.327403	0.436924	0.098284	0.249232	0.174727
5	3000	0.622185	0.08078	0.243452	0.322782	–	–	–	–								
10	10	0.3597	0.0472	0.176	0.0781	0.2746	0.0442	0.1836	0.0560	0.437768	0.045415	0.276732	0.075009	0.358406	0.046993	0.278035	0.059429
10	20	0.3874	0.0478	0.2124	0.0790	0.3015	0.0461	0.2142	0.0583	0.44349	0.048928	0.284487	0.08088	0.362884	0.053646	0.283244	0.067895
10	60	0.3884	0.0438	0.2137	0.0725	0.3033	0.0465	0.2162	0.0588	0.438248	0.047212	0.277747	0.078045	0.355298	0.047748	0.27471	0.06043
10	100	0.422	0.0471	0.2569	0.0778	0.3392	0.0461	0.2566	0.0583	0.441988	0.05029	0.282557	0.083133	0.361566	0.052904	0.281762	0.066957
10	300	0.4087	0.0525	0.2375	0.0874	0.3259	0.0527	0.2406	0.0669	0.425279	0.047684	0.261072	0.078824	0.346883	0.048961	0.265243	0.061966
10	3000	0.404174	0.052263	0.232612	0.086273	–	–	–	–								
20	10	0.2381	0.0197	0.1479	0.0248	0.1946	0.0191	0.1495	0.0213	0.361286	0.024374	0.285971	0.030439	0.318957	0.026121	0.281038	0.029101
20	20	0.2858	0.0261	0.2016	0.0326	0.2446	0.0248	0.2026	0.0276	0.359754	0.024834	0.284431	0.031021	0.322145	0.024238	0.284486	0.027006
20	60	0.3413	0.0263	0.2638	0.0329	0.2943	0.0265	0.2551	0.0296	0.349958	0.027155	0.273483	0.03392	0.316207	0.028292	0.278219	0.031523
20	100	0.3258	0.0268	0.2464	0.0335	0.2895	0.0278	0.2500	0.0310	0.367539	0.02596	0.293132	0.032428	0.32772	0.025885	0.290371	0.028841
20	300	0.3642	0.0263	0.2889	0.0329	0.3226	0.0260	0.2847	0.0290	0.360397	0.025708	0.28515	0.032113	0.321492	0.02609	0.283797	0.02907
20	3000	0.347561	0.024317	0.270196	0.030389	–	–	–	–								
30	10	0.2116	0.0142	0.1529	0.0164	0.1862	0.0142	0.1569	0.0153	0.340816	0.019379	0.291909	0.022347	0.312059	0.020682	0.287451	0.022181
30	20	0.2712	0.0158	0.2171	0.0183	0.2424	0.0158	0.2153	0.0169	0.339644	0.017311	0.290729	0.019971	0.316834	0.018024	0.292435	0.019334
30	60	0.3173	0.0189	0.2667	0.0218	0.2880	0.0179	0.2626	0.0192	0.337467	0.019007	0.28839	0.021927	0.311152	0.019402	0.286551	0.020813
30	100	0.3157	0.0173	0.265	0.0199	0.2898	0.0176	0.2644	0.0189	0.339717	0.016295	0.290807	0.018798	0.310877	0.017103	0.286266	0.018346
30	300	0.3354	0.0174	0.2859	0.0201	0.3107	0.0188	0.2860	0.0202	0.34138	0.016921	0.292593	0.01952	0.312447	0.017927	0.287891	0.01923
30	3000	0.318702	0.02053	0.267844	0.023724	–	–	–	–	–	–	–	–	–	–	–	–

Note: R^2_{ht} is the information theory R^2 measure of the full model; R^2_{adj} is the adjusted coefficient of determination of the full model; $R^2_{\text{ht.R}}$ is the information theory R^2 measure of the reduced model; $R^2_{\text{adj.R}}$is the adjusted coefficient of determination of the reduced model.

Abbreviation: IQR, interquartile range.

10 and one might expect overfitting to have an influence. Similarly there are 30 trials contributing to the penalised likelihood approach, here one might expect the dominant issues to be inefficiency and underestimation. When we look instead at the mean R^2_{adjC}, which at least removes the issue of overfitting and inflation of R^2, we see that in fact the penalised likelihood technique, $R^2_{\text{adjC}} = 0.481$, is out performing the trial removal technique, $R^2_{\text{adjC}} = 0.367$. To obtain a clearer comparison of the techniques, we turn our attention instead to the R^2_{adjC} results.

In the case of strong surrogacy as presented in Table 2 the R^2_{adjC} for the penalised likelihood approach shows less bias and more precision of estimates in all settings that have 60 participants per trial or fewer. The benefits of the penalised likelihood approach increase as the number and size of the trials reduces.

In the case of weak surrogacy, presented in Table 3, the R^2_{adjC} for the penalised likelihood approach displays less bias and more precision in all settings. As in the strong surrogacy setting, the benefit of the penalised likelihood approach increases as the number and size of the trials decreases.

In Table 4 the results in the binary-ordinal case underestimate trial level surrogacy in comparison to the continuous-continuous case and show less precision in all settings. This underestimation worsens as the number of trials increases and although particularly bad for small numbers of patients per trial it can even be seen in the case of 300 patients per trial. If we look at the additional simulation for 3000 patients per trial, we see that the results are much closer to the true value of 0.90, and more in line with that seen for the continuous case with 300 patients per trial. This is evidence that inefficiency is the cause of the bias.

To investigate the overestimation seen predominately for weak surrogacy and small numbers of trials we focus first on where $R^2_{\text{ht}} = 0.30$. See that R^2_{ht} is inflated even when the size of the trials is large. The inflation of the results is particularly bad when the number of trials is small, such that the surrogate might erroneously appear to be moderately good. The reduced model, $R^2_{\text{ht.R}}$, removes some but not all of the inflation of the R^2_{ht}. For instance, for 5 trials and 300 participants the mean $R^2_{\text{ht}} = 0.600$, $R^2_{\text{adjC}} = 0.198$, $R^2_{\text{ht.R}} = 0.407$ and $R^2_{\text{adjC.R}} = 0.207$. The estimate based on the reduced model, $R^2_{\text{ht.R}}$, is still large compared to the true value of 0.30 although not as poor as the full model R^2_{ht} estimate. Under both the full and reduced models the R^2_{adjC} appears to remove all the overestimation in results.

When surrogacy is strong, see Table 4, the inflation of R^2_{ht} is noticeable in comparison to the R^2_{adjC} (mostly where the number of trials is small) but not as severe as in the weak surrogacy setting. Again the $R^2_{\text{ht.R}}$ gives less biased estimation than the R^2_{ht} but not as good as the R^2_{adjC} in regard to overfitting in the smaller trial sizes and trial numbers.

We also compared weak surrogacy for the binary-ordinal to the continuous-continuous setting, see Table 5. Overestimation was seen to behave similarly in both settings in respect to the values of R^2_{ht} and the comparative advantages of R^2_{ht}, R^2_{adjC} and $R^2_{\text{ht.R}}$ described above.

Since the adjusted models remove issues of bias due to overfitting we can compare R^2_{adjC} versus $R^2_{\text{adjC.R}}$ to see clearly the benefits of the full versus reduced models. In all the continuous-continuous settings and the discrete setting where the strength of surrogacy is weak the full model shows little benefit. However in the case of strong surrogacy, see Table 4, R^2_{adjC} is very marginally but consistently less biased than the $R^2_{\text{adjC.R}}$. Presumably this is due to the trial intercept variable including a small amount of information that helps counteract some of the inefficiency present in the discrete setting. However, conversely, in all settings continuous or discrete the precision of the $R^2_{\text{adjC.R}}$ is slightly better than the R^2_{adjC} , increasingly so as the number of trials decreases.

Table S1 outlines the percentage of separation in the various settings: as might be expected the rates were similar regardless of the number of trials. Quasi-complete separation in the binary case was ~60% for 10 participants per trial falling to ~5% for 300 participants per trial. Complete separation only occurred in 2% of trials where there were 10 participants per trial and 0% in all other settings. Separation in ordinal outcomes is always defined as semi-complete[31] and this occurred in ~20% of trials for 10 participants per trial down to ~1% for 60 participants per trial and 0% otherwise.

6 | DISCUSSION

Surrogacy assessment is a complex issue and many statistical approaches have been suggested. An accessible, practically sound and well-developed approach is based on information theory. This approach assesses surrogacy at both the individual patient and trial levels. At the trial level and in the context of discrete outcomes we have identified three issues concerning the information theory approach that have not previously been recognised or investigated.

The first of these is underestimation, which counter-intuitively increases as the number of trials increases and is worse where there are few patients compared to the number of trials. We demonstrated through simulation and investigation of the wider literature that this was due to inefficiency of the two-stage nature of the approach and the use of uninformative discrete outcomes. Continuous outcomes by comparison were minimally affected.

The second issue of bias is that of overestimation. This was worse for small numbers of trials and a weak level of surrogacy. The overestimation was because of overfitting in the second stage of modelling, due to having too few data points available (one per trial) where the number of trials is small. The simulated results for five trials and a weak surrogate are such that a poor surrogate under investigation may erroneously appear to be moderately good. We showed that overfitting of weak level surrogacy was also present for continuous outcomes, supporting our hypothesis that this issue applies across all types of outcome.

Previously, researchers have suggested that a 'full' model should be used instead of a 'reduced' model with fewer explanatory variables. Our simulations show that in fact the reduced models give less biased estimation. This is certainly true in the presence of overfitting and in general it is hard to see much if any benefit of the full model. We also showed that an adjusted R^2 based on the classic coefficient of determination removed issues of overfitting in estimation. A comparison of the adjusted versus unadjusted R^2 shows the large impact overfitting is having on the results and allows us to clearly attribute this bias to overfitting. Based on these simulations, we would advise that surrogacy assessments are based on reduced models, alongside a check of the adjusted R^2 for the reduced model to make sure that the results are not overly optimistic (especially where the number of participants or numbers of trials are small).

Finally, we outlined how poor estimation of treatment effects in the presence of separation at the first stage of trial level surrogacy evaluation leads to biased estimation of the level of surrogacy. We proposed the penalised likelihood technique of Firth[19] as a solution to this. Under simulation investigation, we found an alternative method—removal of trials containing separation from the evaluation—resulted in a large amount of discarded data. This demonstrated how frequently separation can impact on the information available from a trial; given the value of this information, identifying a solution to this issue was critically important. The penalised likelihood technique provides improved estimation and precision without the loss of precious trial information. While this benefit was greater where the number and size of trials is small, these are realistic settings for surrogacy assessments. This technique provides a practical and effective solution to the pervasive issue of separation when assessing surrogacy using the information theory approach for discrete outcomes, and can be easily adopted for any combination of binary or ordinal surrogate and true outcomes. Furthermore, our work indicates that this technique could improve analysis of discrete outcomes in clinical trials research more generally where sparse data is an issue. We developed the command FixedDiscrDiscrIT in the R package Surrogate[32] using the above methodology, allowing practical application of information theoretic surrogacy evaluations in the presence of sparse data.

The use of unvalidated surrogates is an issue that is much debated in clinical trials research. Efforts to encourage researchers to adopt statistical approaches to surrogacy evaluation continue and are only strengthened through refinement of the approaches available. The information theory approach has been centre stage during this undertaking. Our work will further underpin this approach through better understanding of its application in practice and resolution of the issues caused by separation in discrete outcomes.

ACKNOWLEDGEMENTS

We would like to thank Cat Graham and Professor Martin Dennis for their advice and permission to use the CLOTs3[20] trial data. H.E. was supported in this work by a Medical Research Council Trials Methodology Research Studentship and C.J.W. by NHS Lothian via the Edinburgh Clinical Trials Unit.

CONFLICT OF INTEREST

The authors declare they have no conflicts of interest.

DATA AVAILABILITY STATEMENT

Research data from the Clots3 trial are not shared.

ORCID

Hannah Ensor https://orcid.org/0000-0003-3052-7287
Christopher J. Weir https://orcid.org/0000-0002-6494-4903

REFERENCES

1. Alonso A, Molenberghs G. Surrogate marker evaluation from an information theory perspective. *Biometrics*. 2006;63(1):180-186.
2. Ensor H, Lee RJ, Sudlow C, Weir CJ. Statistical approaches for evaluating surrogate outcomes in clinical trials: a systematic review. *J Biopharm Stat*. 2016;26(5):859-879.

3. Tilahun A, Pryseley A, Alonso A, Molenberghs G. Flexible surrogate marker evaluation from several randomized clinical trials with continuous endpoints, using R and SAS. *Comput Stat Data Anal*. 2007;51(9):4152-4163.
4. Tilahun A, Pryseley A, Alonso A, Molenberghs G. Information theory–based surrogate marker evaluation from several randomized clinical trials with binary endpoints, using SAS. *J Biopharm Stat*. 2008;18(2):326-341.
5. Ensor H, Weir CJ. Evaluation of surrogacy in the multi-trial setting based on information theory: an extension to ordinal outcomes. *J Biopharm Stat*. 2020;30(2):364-376.
6. Pryseley A, Tilahun A, Alonso A, Molenberghs G. An information-theoretic approach to surrogate-marker evaluation with failure time endpoints. *Lifetime Data Anal*. 2011;17(2):195-214.
7. Pryseley A, Tilahun A, Alonso A, Molenberghs G. Using earlier measures in a longitudinal sequence as a potential surrogate for a later one. *Comput Stat Data Anal*. 2010;54(5):1342-1354. https://doi.org/10.1016/j.csda.2009.11.024
8. Shi Q, Sargent DJ. Meta-analysis for the evaluation of surrogate endpoints in cancer clinical trials. *Int J Clin Oncol*. 2009;14(2):102-111.
9. Kemp R, Prasad V. Surrogate endpoints in oncology: when are they acceptable for regulatory and clinical decisions, and are they currently overused? *BMC Med*. 2017;15(1):134. https://doi.org/10.1186/s12916-017-0902-9
10. Ciani O, Buyse M, Drummond M, Rasi G, Saad ED, Taylor RS. Use of surrogate end points in healthcare policy: a proposal for adoption of a validation framework. *Nat Rev Drug Discov*. 2016;15(7):516-516.
11. Ciani O, Buyse M, Drummond M, Rasi G, Saad ED, Taylor RS. Time to review the role of surrogate end points in health policy: state of the art and the way forward. *Value Health*. 2017;20(3):487-495.
12. Alonso A, Bigirumurame T, Burzykowski T, et al. *Applied Surrogate Endpoint Evaluation Methods with SAS and R*. New York: CRC Press; 2016.
13. Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H. The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics*. 2000;1(1):49-67.
14. Tibaldi F, Abrahantes JC, Molenberghs G, et al. Simplified hierarchical linear models for the evaluation of surrogate endpoints. *J Stat Comput Simulat*. 2003;73(9):643-658.
15. Shannon CE, Weaver W. *A Mathematical Theory of Communication*. New York: The Bell system technical journal; 1948.
16. Alonso A, Molenberghs G, Geys H, et al. A unifying approach for surrogate marker validation based on Prentice's criteria. *Stat Med*. 2005;25(2):205-221.
17. Kent JT. Information gain and a general measure of correlation. *Biometrika*. 1983;70(1):163-173.
18. SAS Global Forum. Convergence Failures in Logistic Regression. Citeseer; 2008.
19. Firth D. Bias reduction of maximum likelihood estimates. *Biometrika*. 1993;80(1):27-38.
20. Dennis M, Sandercock P, Graham C, et al. The Clots in Legs Or sTockings after Stroke (CLOTS) 3 trial: a randomised controlled trial to determine whether or not intermittent pneumatic compression reduces the risk of post-stroke deep vein thrombosis and to estimate its cost-effectiveness. *Health Technol Assess*. 2015;19(76):1.
21. Abrahantes JC, Molenberghs G, Burzykowski T, Shkedy Z, Abad AA, Renard D. Choice of units of analysis and modeling strategies in multilevel hierarchical models. *Comput Stat Data Anal*. 2004;47(3):537-563.
22. Heinze G, Schemper M. A solution to the problem of separation in logistic regression. *Stat Med*. 2002;21(16):2409-2419.
23. Jeffreys H. An invariant form for the prior probability in estimation problems. *Proc R Soc Lond A Math Phys Sci*. 1946;186:453-461.
24. OrdinalLogisticBiplot. Biplot representations of ordinal variables [program]; 2013.
25. Molenberghs G, Verbeke G, Iddi S. Pseudo-likelihood methodology for partitioned large and complex samples. *Stat Prob Lett*. 2011;81(7):892-901.
26. Taylor AB, West SG, Aiken LS. Loss of power in logistic, ordinal logistic, and probit regression when an outcome variable is coarsely categorized. *Educ Psychol Meas*. 2006;66(2):228-239.
27. Taylor JM, Yu M. Bias and efficiency loss due to categorizing an explanatory variable. *J Multivar Anal*. 2002;83(1):248-263.
28. Nagelkerke NJ. A note on a general definition of the coefficient of determination. *Biometrika*. 1991;78(3):691-692.
29. Pryseley A, Tilahun A, Alonso A, Molenberghs G. Information-theory based surrogate marker evaluation from several randomized clinical trials with continuous true and binary surrogate endpoints. *Clin Trials*. 2007;4(6):587-597.
30. Burzykowski T, Molenberghs G, Buyse M. *The Evaluation of Surrogate Endpoints*. New York: Springer; 2005.
31. Agresti A. *Categorical Data Analysis*. Hoboken, New Jersey: John Wiley & Sons; 2014.
32. Wim Van der Elst PM, Poveda AF, Alonso A, Ensor HM, Molenberghs CJWG. Package surrogate: evaluation of surrogate endpoints in clinical trials. R Package; 2020

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Ensor H, Weir CJ. Separation and the information theory surrogate evaluation approach: A penalised likelihood solution. *Pharmaceutical Statistics*. 2021;1-14. https://doi.org/10.1002/pst.2152