


Evaluation of surrogacy in the multi-trial setting based on information theory: an extension to ordinal outcomes

Hannah Ensor  and Christopher J. Weir

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

ABSTRACT

Summary: In clinical trials, surrogate outcomes are early measures of treatment effect that are used to predict treatment effect on a later primary outcome of interest: the primary outcome therefore does not need to be observed and trials can be shortened. Evaluating surrogates is a complex area as a given treatment can act through multiple pathways, some of which may circumvent the surrogate. One of the best established and practically sound approaches to surrogacy evaluation is based on information theory. We have extended this approach to the case of ordinal outcomes, which are used as primary outcomes in many medical areas. This extension provides researchers with the means of evaluating surrogates in this setting, which expands the usefulness of the information theory approach while also demonstrating its versatility.

ARTICLE HISTORY

Received 12 December 2017

Accepted 18 November 2019

KEYWORDS

Clinical trials; information theory; surrogacy evaluation

1. Introduction


It is legitimate to use a surrogate in place of the true or primary outcome of interest in a clinical trial if it can be established that it informs on the treatment effect on the true outcome. In so doing, a clinical trial can be conducted with a smaller sample size and efficacious treatments can be made available to patients in a more timely fashion. However, the confidence placed in a “legitimate” surrogate can only be as strong as the means of establishing its validity. Baker and Kramer (2003) stated that where treatments work through multiple pathways (as is often the case) surrogacy assessment is difficult. Many different approaches for the evaluation of surrogates have been suggested (Alonso and Molenberghs 2007; Frangakis and Rubin 2004; Molenberghs et al. 2008; Robins and Greenland 1992). For a systematic review of methods see Ensor et al. (2016). These tend to examine whether surrogates are informative at both the individual patient level and the clinical trial level.

The work presented here aims to extend multi-trial information theory-based surrogate evaluation to the case of ordinal outcomes. In so doing we allow researchers to evaluate surrogates in areas where ordinal outcomes are used, for instance in stroke where the Oxford Handicap Scale (Bamford et al. 1989) is often measured.

Various existing surrogate evaluation approaches could be extended to the case of ordinal outcomes, including the direct and indirect effects, principal stratification and information theory approaches (Alonso and Molenberghs 2007; Frangakis and Rubin 2004; Robins and Greenland 1992). Aside from quantitatively evaluating the potential surrogate, we would wish any such approach to have four main properties. An approach should be: (a) practically viable; (b) able to inform on the causal nature of relationships between the surrogate and true outcome; (c) able to identify the surrogate paradox. The surrogate paradox occurs when there are positive treatment

CONTACT Hannah Ensor  Hannah.Ensor@ed.ac.uk  Edinburgh Clinical Trials Unit, Usher Institute, University of Edinburgh, Nine Bioquarter, 9 Little France Road, Edinburgh, EH16 4UX, United Kingdom

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lbps.

 Supplemental data for this article can be accessed on the [publisher's website](#).

© 2019 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

effects on the surrogate, and a positive relationship between the surrogate and true outcome, but a negative treatment effect on the true outcome; and (d) inform on the surrogate's transportability or predictive ability, a fundamental requirement of surrogacy whereby surrogates evaluated in one trial would be able to inform on the treatment effect on the true outcome in a new trial.

Pragmatic multi-trial approaches, including meta-analytical (Buyse et al. 2000) and information theory (Alonso and Molenberghs 2007) are well-established methods that fulfil to a good standard all the above criteria. Therefore, we consider the multi-trial approaches to be the most appropriate for extension to ordinal outcomes. These approaches assess surrogacy at two levels: the individual patient and trial levels. In simple terms, correlation is an insufficient measure of surrogacy because it ignores treatment mechanisms of action and can lead to the surrogate paradox. In multi-trial approaches, the individual patient measure of surrogacy is essentially a correlation but treatment allocation is taken into account. At the trial level, multi-trial approaches provide a measure of the predictive ability of the surrogate to determine whether a surrogate could inform on the likely treatment effect on the primary outcome in a new trial, i.e. its transportability. This satisfies one of the primary aims of a valid surrogate. Combined these measures provide a methodologically sound and practically useful assessment of surrogacy that goes beyond a simple measurement of correlation.

The multi-trial information theory approach has fewer computational and interpretational issues compared to earlier multi-trial approaches and provides consistent interpretation across settings (for example, ordinal or continuous outcomes) (Alonso and Molenberghs 2007). Given these strong methodological and practical advantages, we select the multi-trial information theory approach here for extension to the case of ordinal outcomes.

Most methodology developed for ordinal outcomes has been in early surrogate evaluation measures such as single trial studies (Molenberghs et al. 2001) or the multi-trial (meta-analytical) approach (Burzykowski et al. 2003; Molenberghs et al. 2002; Renard et al. 2002); none have been evaluated via simulation. Under the meta-analytical approach Renard et al. (2002) briefly outline a latent variable approach; Burzykowski et al. (2003) present methodology for an ordinal surrogate and time to event true outcome; and Alonso et al. (2002) investigate the setting where one of the surrogate or true outcome is ordinal and the other continuous. In contrast, our work provides a fully developed methodological extension to the ordinal case in the multi-trial setting, building on the established strengths of the information theory approach to surrogacy evaluation. This has been evaluated by an extensive simulation study incorporating many settings not investigated previously, including weak strengths of surrogacy; discordant strengths of surrogacy at trial and individual levels; ceiling effects for categorical outcomes; as well as an investigation of the impact of non-proportional odds.

In Section 2 we outline the information theory approach and how this can be extended to the case of ordinal outcomes. We cover the “binary-ordinal” setting where the surrogate is binary and the true outcome ordinal; the theory developed could also be applied to the ordinal-ordinal setting with some minor modifications. Section 3 presents a simulation study to evaluate the properties of the ordinal extension. Section 4 illustrates the method using a case study from the stroke clinical trial CLOTS3 (Dennis et al. 2015) and Section 5 discusses our methodology extension in the broader context.

2. Methods

In what follows, the surrogate is denoted S , treatment is Z and the true outcome is T . There are $i = 1, 2, \dots, N$ trials, and $j = 1, 2, \dots, n_i$ patients per trial. $N_T = \sum_i n_i$ is the total number of patients in all trials. The ordinal true outcome has W ordered categories.

2.1. The information theory approach

Alonso and Molenberghs (2007) proposed an information theory surrogate evaluation measure based on the concepts of entropy and information theory by Shannon (1948). Information theory uses the central concept of entropy to measure the “information, choice and uncertainty” in a random variable.

In the discrete case, entropy can be represented 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, \dots, k_{m_y} and probabilities p_1, p_2, \dots, p_{m_y} respectively. Conditional, $H(Y|X)$, and joint entropy, $H(Y, X)$, can be straightforwardly defined. And differential entropy measures information in the continuous case, $h_d(Y) = -\int_{-\infty}^{\infty} f_Y(y) \log\{f_Y(y)\} dy$.

A concept of fundamental importance is 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. Another useful concept for comparing random variables is the entropy power, obtained by maximising the entropy of a continuous random variable, defined as $EP(Y) = \frac{1}{(2\pi e)} e^{2h(Y)}$. See Shannon (1948) for a full list of the properties of entropy and the mutual information.

These concepts are useful in surrogate evaluation as, at the individual level, we are interested in the amount of information on T (or ‘treatment effects on T ’ at the trial level) covered by our knowledge of S (or ‘treatment effects on S ’ at the trial level).

2.1.1. Individual level: information theory approach

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

$$R_h^2 = \frac{EP(T) - EP(T|S)}{EP(T)} \tag{1}$$

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: 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 ; and $R_h^2 = 0$ if and only if T and S are independent.

Alonso and Molenberghs (2007) suggested a multi-trial framework R_h^2 ; as shown in Equation (2) to enable transportability of results for the information theory approach.

$$R_h^2 = \sum_{i=1}^{N_q} \vartheta_i R_{hi}^2 = 1 - \sum_{i=1}^{N_q} \vartheta_i e^{-2I_i(S_i, T_i)},$$

where

$$\vartheta_i > 0 \forall i, \sum_{i=1}^{N_q} \vartheta_i = 1 \tag{2}$$

For N trials there are N_q possible values of R_h^2 , the R_h^2 for the i^{th} trial since trials can be clustered depending, say, on q different characteristics (e.g. centre, country, treating physician). There are many different choices for the set of unknown weights, ϑ_i in (2). The choice of which leads to an uncountable set of parameters, Ω_h , each parameter of which could act as a single meaningful measure of R_h^2 in the multi-trial setting:

$$\Omega_h = \left\{ \Phi_h : \Phi_h = 1 - \sum_{i=1}^{N_q} \vartheta_i e_i^{-2I_i(S_i, T_i)}, \text{ where } \vartheta_i > 0 \forall i, \sum_{i=1}^{N_q} \vartheta_i = 1 \right\} \tag{3}$$

where Φ_h are the parameters of the set Ω_h . Alonso and Molenberghs (2007) highlighted the likelihood reduction factor (LRF) as a good candidate from Ω_h which provides a useful route to defining $\sum_{i=1}^{N_q} \vartheta_i$. The LRF is a measure of information gain that has been considered under an information theory framework by several authors (Brillinger 2004; Joe 1989; Kullback 1997; Linfoot 1957).

The LRF is particularly useful for surrogacy evaluation as it ranges in the unit interval and has a consistent interpretation across settings: this is a key point as previous approaches could not provide this. Furthermore, it is possible that a high-dimension integral would be needed in the calculation of $I(T,S)$ which the LRF avoids, and as we will expand on in section 2.1.1.1 the LRF provides consistent estimation of R_h^2 (Alonso et al. 2016; Alonso and Molenberghs 2007). Finally,

previous approaches to surrogacy assessment relied on computationally intensive joint models of S and T, but the LRF assesses just the conditional model of T|S and the marginal model of T and hence avoids this issue.

2.1.1.1. LRF at the individual level: the continuous setting. The LRF was proposed by Alonso et al. (2006) based on the ideas of Kent (1983). At the individual level, the LRF is based on the amount of information gained about the true outcome after accounting for the surrogate which was proposed as a general measure of correlation. Alonso et al. (2005) proposed modelling (4) and (5) for each trial i (linear models are presented here, whereas generalised linear models were originally given):

$$T_{ij} = \mu_i + \beta_i Z_{ij} + \varepsilon_{T_{ij}} \tag{4}$$

$$T_i = \theta_{0_i} + \theta_{1_i} Z_{ij} + \theta_{2_i} S_{ij} + \varepsilon_{T|S_{ij}} \tag{5}$$

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 amount of information on the true outcome gained from the surrogate is calculated via the difference in the log-likelihood between (4) and (5) which is formally expressed as G_i^2 , for each trial i . LL_0 is the log-likelihood for the unsaturated model, in this case (4), and LL_1 for the saturated model, (5), for trial i . $G_i^2 = 2(LL_1 - LL_0)$.

The LRF is then calculated:

$$LRF = 1 - \frac{1}{N} \sum_{i=1}^N \exp\left(-\frac{G_i^2}{n_i}\right) \tag{6}$$

To demonstrate the link between the LRF and R_h^2 , consider the i^{th} trial and joint density function $f(T_i, S_i | \theta_i)$ of (T_i, S_i) , where we have $\theta_i = (\theta_{*i}, \theta_{2i})$. Where θ_{2i} represents the dependence between S and T, $\hat{\theta}_{*i}$ is the maximum likelihood estimator under the null hypothesis of independence ($\theta_{2i} = 0$), and $\hat{\theta}_i$ is the maximum likelihood estimator for the saturated model. We can express $\frac{1}{n_i} G_i^2 = \frac{1}{n_i} \sum_i [\log\{f(t_i | s_i, \hat{\theta}_i)\} - \log\{f(t_i | \hat{\theta}_{*i})\}]$. If $\hat{\theta}_{*i}$ converges to θ_{*i} in probability then $\frac{1}{n_i} G_i^2 \rightarrow I_i(S_i, T_i)$ under general regularity conditions, hence $\frac{1}{n_i} G_i^2$ is a consistent estimator of $I_i(S_i, T_i)$.

Using the estimator $\frac{1}{n_i} G_i^2$ of $I_i(S_i, T_i)$ we have $LRF = \hat{R}_h^2 = 1 - \sum_{i=1}^N \frac{1}{N} e^{-\frac{1}{n_i} G_i^2}$, which is a special case of R_h^2 in (3) where $\vartheta_i = \frac{1}{N}$. Therefore, the LRF is a consistent estimator of R_h^2 (Alonso and Molenberghs 2007; Brillinger 2004). For a full proof see the supplementary material of (Alonso and Molenberghs 2007).

2.1.1.2. LRF at the individual level: extension to the binary-ordinal setting. The LRF can be used to calculate R_h^2 for a binary surrogate and ordinal true outcome. At the individual level, the LRF can be applied in the binary-ordinal setting in the same manner as in the continuous case using (6), based in this case on the difference $G^2 = 2(LL_1 - LL_0)$ of the following proportional odds models:

$$\text{logit}\{P(T_{ij} \leq w)\} = \mu_{T_{w_i}} + \beta_i Z_{ij} \tag{7}$$

$$\text{logit}\{P(T_{ij} \leq w)\} = \theta_{0_{w_i}} + \theta_{1_i} Z_{ij} + \theta_{2_i} S_{ij} \tag{8}$$

where $w = 1, \dots, W - 1$, and W is the number of categories in the ordinal true outcome. For trial i , $\mu_{T_{w_i}}$ and $\theta_{0_{w_i}}$ are intercept parameters for each cut point of the ordinal true outcome, β_i and θ_{1_i} represent the treatment effect on the true outcome and θ_{2_i} is the surrogate parameter. Again, the LRF is based on the amount of information gained on the true outcome after adjusting for 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 (Kent 1983). Alonso and Molenberghs (2007) showed that $R_h^2 \leq 1 - e^{-2H(T)}$, where $H(T)$ represents the entropy of T. They also suggested that $H(T)$ can be approximated based on the log-likelihood of the intercept-only model of true outcome ($\text{logit}\{P(T_{ij} \leq w)\} = \theta_3$, where θ_3 is the intercept parameter). Alonso and Molenberghs (2007) therefore proposed rescaling R_h^2 as calculated by the LRF in (6) by:

$$\widehat{R}_h^2 = \frac{R_h^2}{1 - e^{-2H(T)}} \tag{9}$$

The LRF thus gives a consistent interpretation at the individual level for both the binary-ordinal and continuous settings.

2.1.2. Trial level: information theory approach

At the trial level, interest is in the relationship between treatment effects on the surrogate and treatment effects on the true outcome. Alonso and Molenberghs (2007) proposed a two-stage approach. At the first stage, the treatment effects for each trial on the surrogate and true outcome are obtained, α_i and β_i respectively. This is done by regressing the surrogate and true outcome on treatment in separate models:

$$S_{ij} = \mu_{S_i} + \alpha_i Z_{ij} + \varepsilon_{S_{ij}} \tag{10}$$

$$T_{ij} = \mu_{T_i} + \beta_i Z_{ij} + \varepsilon_{T_{ij}} \tag{11}$$

where μ_{S_i}, μ_{T_i} , represent the mean intercept and α_i, β_i the treatment effects for S and T, respectively.

Using the treatment effect estimates for S and T from these models, $\hat{\alpha}_i$ and $\hat{\beta}_i$ respectively, we calculate the information theory surrogacy measure R_{ht}^2 , where the subscript t indicates that we are now considering trial-level surrogacy, through:

$$R_{ht}^2 = \frac{EP(\hat{\beta}) - EP(\hat{\beta}|\hat{\alpha})}{EP(\hat{\beta})} \tag{12}$$

where $EP(\hat{\beta})$ is the entropy power of the distribution of treatment effect estimates on T across the i trials and $EP(\hat{\beta}|\hat{\alpha})$ is the entropy power of the distribution of treatment effect estimates on T given those on S. R_{ht}^2 can be interpreted as the amount of uncertainty in the treatment effect on T removed through knowledge of the treatment effect on S.

2.1.2.1. The LRF at the trial level: the continuous setting. The LRF can be applied to calculate R_{ht}^2 in the continuous-continuous case. In order to do this (10) and (11) are again modelled to obtain treatment estimates $\hat{\alpha}_i$ and $\hat{\beta}_i$ and $\hat{\mu}_{S_i}$. At the second stage, two further models of the treatment effect on the true outcome are required:

$$\hat{\beta}_i = \gamma_3 + \varepsilon_{\beta_i} \tag{13}$$

$$\hat{\beta}_i = \gamma_0 + \gamma_1 \hat{\mu}_{S_i} + \gamma_2 \hat{\alpha}_i + \varepsilon_{\beta|\mu, \alpha_i} \tag{14}$$

where γ_3 and γ_0 are the intercept parameters with and without adjustment for the surrogate treatment effects and γ_1 and γ_2 are the parameters for the surrogate intercept and treatment effect estimates provided from stage one. The difference in log-likelihood between these two models can then be calculated and the LRF applied as in (15).

$$LRF = \widehat{R}_{ht}^2 = 1 - \exp\left(-\frac{G_2}{N}\right) \tag{15}$$

In a similar fashion to the LRF at the individual level, it can be shown that the LRF is a consistent estimator of R_{ht}^2 (Alonso et al. 2016).

2.1.2.2. The LRF at the trial level: extension to the binary-ordinal setting. In the binary-ordinal setting, the key difference in the approach is in the models used at the first stage. Here a generalised linear and proportional odds model are required for the surrogate and true outcome, respectively:

$$\text{logit}\{P(S_{ij} = 1)\} = \mu_{S_i} + \alpha_i Z_{ij} \tag{16}$$

$$\text{logit}\{P(T_{ij} \leq w)\} = \mu_{T_{w_i}} + \beta_i Z_{ij} \tag{17}$$

where $w = 1, \dots, W - 1$, and W is the number of categories in the ordinal true outcome, $\mu_{T_{w_i}}$ is the set of intercept parameters for each of the $W-1$ cut points of the ordinal true outcome and all other parameters are analogous to the continuous case. The second stage models (13) and (14) can be fitted in the same manner as in the continuous setting using the parameters of (16) and (17), and the LRF applied as in (15). The LRF has a consistent interpretation at the trial level for the continuous-continuous and binary-ordinal settings, and it can easily be seen how this would be the case for other settings.

2.2. Confidence intervals – all settings

A confidence interval based on the non-central χ^2 distribution for R_{ht}^2 may be calculated as per (Kent 1983):

$$\left\{ 1 - \exp\left(-\frac{\gamma_{1:\alpha/2}(G^2)}{N}\right), 1 - \exp\left(-\frac{\delta_{1:\alpha/2}(G^2)}{N}\right) \right\}$$

where $\gamma_{1:\alpha}$ and $\delta_{1:\alpha}$ are defined by $P[\chi_1^2\{\gamma_{1:\alpha}(G^2)\} \geq G^2] = \alpha$ and $P[\chi_1^2\{\delta_{1:\alpha}(G^2)\} \leq G^2] = \alpha$, and χ_1^2 represents the non-central chi-squared distribution with 1 degree of freedom. The above is true unless $P\{\chi_1^2(0) \geq G^2\} > \alpha$ in which case $\gamma_{1:\alpha}(G^2) = 0$.

R_h^2 on the other hand has multiple G_i^2 . Previous publications have computed non-parametric bootstrap confidence intervals in this setting and we follow that methodology (Alonso et al. 2006).

3. Simulation study

3.1. Set-up

The practical worth of the approach is demonstrated via a thorough simulation study using R, based on the approach of (Tilahun et al. 2008). Different scenarios were simulated to see how the R^2 measures perform when different numbers of trials and sizes of trial are available. We reported the median point estimate and median upper and lower confidence limits over 250 simulations for each scenario investigated. We use the methodology of the precursor to the information theory approach, the meta-analytical approach, to set up the simulation as conducted by many previous authors Tilahun et al. (2008). The normal joint mixed model (17) gives the basis for the data generation:

$$S_{ij} = \mu_S + m_{S_i} + \alpha Z_{ij} + a_i Z_{ij} + \varepsilon_{S_{ij}} \tag{18}$$

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

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. $(\varepsilon_{S_{ij}}, \varepsilon_{T_{ij}}) \sim N(0, \Sigma)$ and random effects, $(m_{S_i}, m_{T_i}, a_i, b_i)^T \sim N(0, D)$, where:

$$D = 3 \begin{pmatrix} 1 & 0.75 & 0 & 0 \\ 0.75 & 1 & 0 & 0 \\ 0 & 0 & 1 & \rho \\ 0 & 0 & \rho & 1 \end{pmatrix}, \text{ where } R_{ht}^2 = \rho^2,$$

$$\Sigma = 3 \begin{pmatrix} 1 & \psi \\ \psi & 1 \end{pmatrix}, \text{ where } R_h^2 = \psi^2.$$

Specific values of D and Σ were chosen in line with Tilahun et al. (2008) as were individual trial intercept and treatment parameters for S and T which were set to $\mu_s = 0.50$, $\mu_T = 0.45$, $\alpha = 0.05$, and $\beta = 0.03$. Their values do not influence the true strength of surrogacy.

Four surrogacy scenarios were simulated: strong, with $R_{ht}^2 = \rho^2 = 0.90$ and $R_h^2 = \psi^2 = 0.64$; weak, with $R_{ht}^2 = \rho^2 = 0.30$ and $R_h^2 = \psi^2 = 0.30$; or to have discordant levels of surrogacy at trial and individual level, $R_{ht}^2 = \rho^2 = 0.90$ and $R_h^2 = \psi^2 = 0.30$; or $R_{ht}^2 = \rho^2 = 0.30$ and $R_h^2 = \psi^2 = 0.64$. After simulating a continuous S and T these were then dichotomised or categorised to represent a binary S and ordinal T. T was set to have seven categories and its distribution was simulated to follow what might be observed in the Oxford Handicap Scale (Van Swieten et al. 1988) investigated in the stroke case study (section 4). We also investigate the setting where the ordinal outcome does not fulfil the proportional odds assumption, by changing for one treatment arm one of the quantiles at which the continuous T is cut to generate the ordinal categorical T. Trial sizes were set to 60, 100, and 300 patients. There were 5, 10, 20 or 30 trials in each simulated data set. There were 250 datasets simulated for each scenario: a total of 15,000 simulations covering all combinations of the strength of surrogacy (4), trial size (3) and number of trials (4) scenarios and in addition the non-proportional odds setting with strong surrogacy for all trial size and number of trials scenarios.

At the individual level in the discrete binary-ordinal case, information theory explores surrogacy at the observed rather than latent scale, and therefore the strength of surrogacy is expected to be lower than on the latent continuous level (Tilahun et al. 2008). This reflects reality, since for example binary measures often represent latent continuous variables and a binary surrogate would be expected to provide less information than a continuous one. Therefore, we expect the maximum surrogacy strength achievable in the observed binary-ordinal setting to be much lower than the ‘true’ strength of surrogacy set at the latent level. We investigated the individual level surrogacy ceiling for a binary surrogate with an ordinal true outcome by further investigating the ideal scenario where $R_{ht}^2 = \rho^2 = 0.90$ and $R_h^2 = \psi^2 = 1$. In this case, 250 data sets were simulated for each scenario: a total of 3,000 simulations covering all combinations of the trial size (3) and number of trials (4) scenarios.

3.2. Results

For strong surrogacy \widehat{R}_h^2 converges to around 0.30 (Table 1) for larger numbers of trials and trial sizes; this is much lower than the 0.64 strength simulated on the latent continuous scale. Equally, for weak surrogacy \widehat{R}_h^2 converges to around 0.13 (Table 2) which is again much lower than the strength of 0.30 simulated on the latent scale. Simulations for the ‘perfect’ surrogate with $R_h^2 = 1$ converge to around $\widehat{R}_h^2 = 0.48$, the ceiling for this binary surrogate for an ordinal true outcome generated from a latent continuous measure, see Table 3.

Unlike individual-level surrogacy, trial-level surrogacy, R_{ht}^2 ought to report the same surrogacy strength at the latent and explicit scales (Tilahun et al. 2008). However, there appears to be some underestimation of \widehat{R}_{ht}^2 for strong surrogacy even where trial sizes are large (Table 1). This is in line with results in the continuous-binary and binary-binary settings (Pryseley et al. 2007; Tilahun et al. 2008). Conversely, where surrogacy is set to be weak (Table 2) there is overestimation of R_{ht}^2 for small trial sizes. Further examination showed this was due to overfitting to the resultant small number of data points (one for each trial) in the regression model used at the second stage of R_{ht}^2 modelling.

Table 1. Simulation study results: Strong surrogacy. True values on the latent continuous scale used to generate data are trial-level surrogacy $R_{ht}^2 = 0.90$, and individual-level surrogacy $R_h^2 = 0.64$ (at the individual level we expect strength of surrogacy in the binary-ordinal setting to be low due to loss of information from moving from continuous to categorical outcomes). 250 simulations were performed for each of the scenarios reported in the table. We present the number and size of trials simulated; the median R of the 250 simulations; median lower and upper limits of the 95% confidence intervals.

Number of trials	Trial size	R_{ht}^2 : Trial-level surrogacy			R_h^2 : Individual-level surrogacy		
		Median \widehat{R}_{ht}^2	Lower 95%CI	Upper 95%CI	Median \widehat{R}_h^2	Lower 95%CI	Upper 95%CI
5	60	0.930	0.404	0.998	0.308	0.228	0.398
5	100	0.934	0.429	0.998	0.307	0.242	0.371
5	300	0.948	0.501	0.999	0.305	0.267	0.343
10	60	0.833	0.349	0.982	0.304	0.245	0.367
10	100	0.847	0.411	0.983	0.298	0.252	0.344
10	300	0.895	0.541	0.989	0.299	0.271	0.325
20	60	0.793	0.454	0.952	0.297	0.257	0.341
20	100	0.826	0.522	0.960	0.300	0.268	0.334
20	300	0.871	0.622	0.970	0.293	0.274	0.312
30	60	0.783	0.512	0.929	0.297	0.263	0.332
30	100	0.823	0.588	0.944	0.296	0.270	0.323
30	300	0.866	0.668	0.958	0.293	0.278	0.310

Table 2. Simulation study results: weak surrogacy. True values on the latent continuous scale used to generate data are trial-level surrogacy $R_{ht}^2 = 0.30$, and individual-level surrogacy $R_h^2 = 0.30$ (at the individual level we expect strength of surrogacy in the binary-ordinal setting to be low due to loss of information from moving from continuous to categorical outcomes). 250 simulations were performed for each of the scenarios reported in the table. We present the number and size of trials simulated; the median R^2 of the 250 simulations; median lower and upper limits of the 95% confidence intervals for the 250 simulations.

Number of trials	Trial size	R_{ht}^2 : Trial-level surrogacy			R_h^2 : Individual-level surrogacy		
		Median \widehat{R}_{ht}^2	Lower 95%CI	Upper 95%CI	Median \widehat{R}_h^2	Lower 95%CI	Upper 95%CI
5	60	0.643	0.028	0.974	0.143	0.086	0.231
5	100	0.682	0.039	0.979	0.140	0.092	0.204
5	300	0.670	0.038	0.977	0.135	0.107	0.171
10	60	0.429	0.012	0.866	0.144	0.104	0.206
10	100	0.393	0.009	0.843	0.137	0.105	0.183
10	300	0.385	0.009	0.832	0.134	0.113	0.158
20	60	0.265	0.010	0.656	0.138	0.114	0.183
20	100	0.303	0.021	0.676	0.136	0.115	0.170
20	300	0.311	0.027	0.673	0.131	0.118	0.149
30	60	0.243	0.019	0.568	0.141	0.122	0.179
30	100	0.271	0.033	0.589	0.136	0.119	0.164
30	300	0.304	0.054	0.610	0.132	0.121	0.147

\widehat{R}_{ht}^2 and \widehat{R}_h^2 estimates where surrogacy strengths differ at trial and individual levels are similar to where surrogacy strengths are consistent, see Table 4. Deviation from the proportional odds assumption also seems to have little impact on results at either level, see Table 5.

4. Case study – CLOTS3

The case study, conducted using data from the randomised trial Clots in Legs Or sTockings after Stroke (CLOTS) 3 trial (Dennis et al. 2015), aimed to determine whether measures taken within 30 days of a stroke could be used as a surrogate in place of death and disability measured 6 months post stroke.

Venous thromboembolism encompasses the ailments: deep vein thrombosis (DVT), a blood clot in the deep veins of the legs; and pulmonary embolism (PE), where clots detach from the veins and

Table 3. Simulation study results: Ceiling effect. True values on the latent continuous scale used to generate data are trial-level surrogacy $R_{ht}^2 = 0.90$, and individual-level surrogacy $R_h^2 = 1$ (at the individual level we expect strength of surrogacy in the binary-ordinal setting to be low due to loss of information from moving from continuous to categorical outcomes). 250 simulations were performed for each of the scenarios reported in the table. We present the number and size of trials simulated; the median R^2 of the 250 simulations; median lower and upper limits of the 95% confidence intervals for the 250 simulations.

Number of trials	Trial size	Individual-level surrogacy		
		Median \widehat{R}_h^2	Lower 95%CI	Upper 95%CI
5	60	0.539	0.398	0.599
5	100	0.548	0.429	0.598
5	300	0.516	0.444	0.569
10	60	0.514	0.419	0.562
10	100	0.510	0.426	0.556
10	300	0.484	0.424	0.540
20	60	0.492	0.425	0.527
20	100	0.500	0.441	0.535
20	300	0.489	0.438	0.520
30	60	0.494	0.438	0.521
30	100	0.488	0.439	0.517
30	300	0.478	0.438	0.508

Table 4. Simulation study results: differing strengths of surrogacy against the case where surrogacy is strong at both levels. 250 simulations were performed for each of the scenarios reported in the table. We present the number and size of trials simulated; and the median R^2 of the 250 simulations. [‡] Both comparisons are between strong level surrogacy at both levels, $R_{ht}^2 = 0.90$ and $R_h^2 = 0.64$, against the case where surrogacy is strong at the level under consideration but weak (either $R_{ht}^2 = 0.30$ or $R_h^2 = 0.30$) at the unreported level. The converse case gives comparable results (results not shown).

Number of trials	Trial size	Median R_{ht}^2 [‡] Trial-level surrogacy		Median R_h^2 [‡] Individual-level surrogacy	
		$R_{ht}^2 = 0.90$ $R_h^2 = 0.64$	$R_{ht}^2 = 0.90$ $R_h^2 = 0.30$	$R_{ht}^2 = 0.90$ $R_h^2 = 0.64$	$R_{ht}^2 = 0.30$ $R_h^2 = 0.64$
5	60	0.930	0.905	0.308	0.303
5	100	0.934	0.934	0.307	0.308
5	300	0.948	0.951	0.305	0.302
10	60	0.833	0.823	0.304	0.294
10	100	0.847	0.851	0.298	0.293
10	300	0.895	0.895	0.299	0.291
20	60	0.793	0.750	0.297	0.292
20	100	0.826	0.811	0.3	0.290
20	300	0.871	0.865	0.293	0.287
30	60	0.783	0.734	0.297	0.291
30	100	0.823	0.803	0.296	0.291
30	300	0.866	0.861	0.293	0.288

cause blockages to the lungs. Venous thromboembolism can be serious enough to cause death or be so debilitating it hinders rehabilitation. Dennis et al. (2013) showed that 20–42% of stroke patients suffer a venous thromboembolism. This result reflects the fact that stroke patients are typically bedbound and often unable to move one side of their body.

A primary measure of ongoing health and survival measured in patients 6 months post stroke is the Oxford Handicap Scale (OHS) (Van Swieten et al. 1988). This is an ordinal measure on a seven-point scale, ranging from no symptoms up to severe disability and death.

CLOTS3 was a 94 centre randomised clinical trial with 2,876 patients. It was conducted to investigate whether intermittent pneumatic compression (IPC) applied to the legs of acute stroke patients reduced the occurrence of DVT (Dennis et al. 2015). CLOTS3 () showed that IPC reduced the odds of DVT by 30 days [OR 0.65 (95% CI 0.51–0.84; $p = .001$) after adjustment for baseline variables] and had a positive impact on survival at 6 months, HR 0.86 (0.74–0.99), $p = .042$.

Table 5. Simulation study results: considering proportional versus non-proportional odds. True values on the latent continuous scale used to generate data are trial-level surrogacy $R_{ht}^2 = 0.90$, and individual-level surrogacy $R_h^2 = 0.64$. 250 simulations were performed for each of the scenarios reported in the table. We present the number and size of trials simulated; and the median R : of the 250 simulations.

Number of trials	Trial size	Median R_{ht}^2		Median R_h^2	
		Trial-level surrogacy		Individual-level surrogacy	
		Proportional	Non-Proportional	Proportional	Non-Proportional
5	60	0.930	0.925	0.308	0.305
5	100	0.934	0.928	0.307	0.308
5	300	0.948	0.947	0.305	0.304
10	60	0.833	0.823	0.304	0.301
10	100	0.847	0.845	0.298	0.292
10	300	0.895	0.890	0.299	0.295
20	60	0.793	0.788	0.297	0.295
20	100	0.826	0.818	0.3	0.295
20	300	0.871	0.866	0.293	0.291
30	60	0.783	0.772	0.297	0.294
30	100	0.823	0.818	0.296	0.293
30	300	0.866	0.862	0.293	0.290

We used this data set to assess whether the occurrence of DVT, PE or death within 30 days is a surrogate for OHS at 6 months. The information theory approach was applied to investigate this and we used study centres in place of trials (Abrahantes et al. 2004). The results shown in Table 6 and Figure 1 indicate that DVT is not a good surrogate for OHS, as \widehat{R}_h^2 is 0.173 95% CI (0.141, 0.188) and \widehat{R}_{ht}^2 is 0.186 95% CI (0.048, 0.374). While there is no established cut-off corresponding to a ‘valid’ surrogate previous publications have suggested that surrogates that exceed 0.80 at both levels can be deemed valid, while if surrogacy strength at either level is below 0.50 surrogacy strength is poor (Alonso et al. 2016). Therefore, these results suggest a poor surrogate.

A sensitivity analysis was conducted to assess whether the potential bias witnessed in the simulation study (where underestimation increased with increased number of trials) may have influenced our results. We regrouped centres so there were fewer groups (results not shown), and we found that while some level of underestimation had taken place the point estimates were still comfortably under 0.50 and therefore our conclusions did not change. Further information on the case study is provided in Appendix A.

5. Discussion

The information theory approach of Alonso and Molenberghs (2007) has previously been extended to failure time outcomes (Pryseley et al. 2011); repeated measures (Alonso et al. 2006); a continuous surrogate and binary true outcome (Pryseley et al. 2007); and binary outcomes (Tilahun et al. 2008). This paper complements these by extending the methodology to the case of a binary surrogate and an ordinal true outcome.

A major strength of the work presented is the wide range of scenarios considered in the simulation study that evaluated the performance of the extended methodology measures of

Table 6. CLOTS3 case study results: Information theory surrogacy estimates for binary DVT surrogate and ordinal OHS true outcome; analysed using a modified information theory approach incorporating a penalized likelihood approach (Firth 1993) to deal with the issue of sparse data.

R_h^2 Individual level	R_{ht}^2 Trial level
0.173	0.186
95% CI (0.141, 0.188))	95% CI (0.048,0.374)

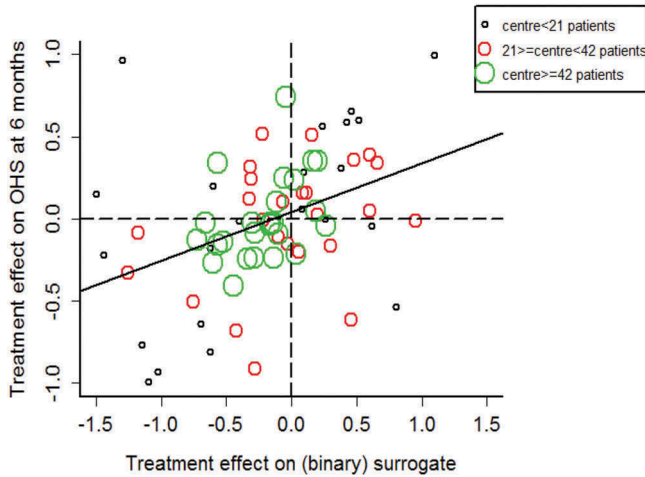


Figure 1. CLOTS3 case study results: Graphical display of information theory surrogacy estimates for binary DVT surrogate and ordinal OHS true outcome; study centre size categorised by the terciles of centre size. The regression line represents the regression of the treatment effects on the true outcome on those for the surrogate. Analysed using a modified information theory approach incorporating a penalized likelihood approach (Firth 1993) to deal with the issue of sparse data.

individual-level surrogacy, R_h^2 , and trial-level surrogacy, R_{ht}^2 , in the binary-ordinal context. Extending previous simulation studies in this area (Tilahun et al. 2007, 2008) we assessed weak strengths of surrogacy, discordant levels of surrogacy at trial and individual levels and investigated the ceiling effect present when using binary and ordinal outcomes. We also completed the first assessment via simulation of the non-proportional odds scenario for ordinal outcomes. A further benefit was the opportunity to provide a clear answer to a question of clinical interest regarding deep vein thrombosis, DVT, as a potential surrogate for long-term outcome following stroke the Oxford Handicap Scale, OHS, using data from the CLOTS (Dennis et al. 2015) randomised controlled trial.

As might have been expected, the simulation study showed that a binary surrogate is less informative than its latent counterpart at the individual level; the ceiling for the binary-ordinal setting is around half the strength of that simulated on the underlying continuum.

Some unexpected underestimation of R_{ht}^2 was observed; we speculate that this is due to inefficiencies in estimation through a combination of the use of a two-stage estimation approach and the involvement of discrete outcomes. Furthermore, overestimation of R_{ht}^2 occurred for weak surrogacy and small numbers of trials, due to overfitting at the second stage of modelling. Assessments of surrogates of this kind might lead researchers to believe incorrectly that they are valid. This is likely to be an issue regardless of the setting (binary-continuous, continuous-continuous etc.) and has not previously been identified. These two sources of bias at the trial level, overfitting and inefficiency, point to some practical issues with the two-stage modelling approach and require further investigation. Deviations from the proportional odds assumption or discordant surrogacy strength at trial and individual levels had little impact on \widehat{R}_{ht}^2 or \widehat{R}_h^2 results with positive implications for the robustness of this surrogacy assessment approach.

In future work, it would be interesting to study the underestimation found in this work in more detail – perhaps in the context of contrasting settings, e.g. time-to-event or repeated measures. Nevertheless, if inefficiency is the root cause of underestimation the discrete case is likely to be the most severely affected. In the discrete outcome setting the issues of estimation in the presence of separation (perfect agreement between two discrete outcomes) is one that might have a large impact on results. Our simulations did not consider small trial sizes where separation is likely to be a substantial issue; however, this important topic should be considered in more detail alongside

potential solutions. Equally, it would be worth establishing if the overfitting witnessed in the case of weak surrogacy is systemic to all settings of the information theory approach.

Overall results from the simulation show that the information theory approach works well in general in the binary-ordinal context, although some issues concerning the two stage nature of the modelling approach for R_{ht}^2 have been identified. Methodologically we have seen that across settings the information theory approach is readily applied and provides a consistent interpretation. The methodological extensions reported here will enable researchers working in clinical areas where ordinal outcomes are important to investigate surrogacy. This work provides further confirmation that information theory is a practical and methodologically sound approach to surrogacy evaluation.

Acknowledgments

We would like to thank Cat Graham and Prof. Martin Dennis for their advice and permission to use the CLOTS3 (Dennis et al. 2013) trial data and Prof. Cathie Sudlow for her insights into the practical application of this work. Christopher Weir was supported in this work by NHS Lothian via the Edinburgh Clinical Trials Unit.

Funding

This work was supported by the Medical Research Council; Rural and Environment Science and Analytical Services Division.

ORCID

Hannah Ensor  <http://orcid.org/0000-0003-3052-7287>

References

- Abrahantes, J. C., G. Molenberghs, T. Burzykowski, Z. Shkedy, A. A. Abad, and D. Renard. 2004. Choice of units of analysis and modeling strategies in multilevel hierarchical models. *Computational Statistics & Data Analysis* 47 (3):537–563. doi:10.1016/j.csda.2003.12.003.
- Alonso, A., and G. Molenberghs. 2007. Surrogate marker evaluation from an information theory perspective. *Biometrics* 63 (1):180–186. doi:10.1111/biom.2007.63.issue-1.
- Alonso, A., G. Molenberghs, H. Geys, M. Buyse, and T. Vangeneugden. 2006. A unifying approach for surrogate marker validation based on Prentice's criteria. *Statistics in Medicine* 25 (2):205–221. doi:10.1002/sim.2315
- Alonso, A., H. Geys, G. Molenberghs, and T. Vangeneugden. 2002. Investigating the criterion validity of psychiatric symptom scales using surrogate marker validation methodology. *Journal of Biopharmaceutical Statistics* 12 (2):161–178. doi:10.1081/BIP-120015741.
- Alonso, A., T. Bigirimurame, T. Burzykowski, M. Buyse, G. Molenberghs, L. Muchene, N. J. Perualila, Z. Shkedy, and W. Van der Elst. 2016. *Applied surrogate endpoint evaluation methods with sas and r*. Boca Raton, FL: CRC Press.
- Baker, S. G., and B. S. Kramer. 2003. A perfect correlate does not a surrogate make. *BMC Medical Research Methodology* 3 (1):16. doi:10.1186/1471-2288-3-16.
- Bamford, J. M., P. A. G. Sandercock, C. P. Warlow, J. Slattery. 1989. Letter to the Editor: Interobserver agreement for the assessment of handicap in stroke patients. *Stroke* 20:828. doi:10.1161/01.STR.20.6.828.
- Brillinger, D. R. 2004. Some data analyses using mutual information. *Brazilian Journal Of Probability and Statistics* 18:163–182.
- Burzykowski, T., G. Molenberghs, and M. Buyse. 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 (1):103–124. doi:10.1111/j.1467-985X.2004.00293.x.
- Buyse, M., G. Molenberghs, T. Burzykowski, D. Renard, and H. Geys. 2000. The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics* 1 (1):49–67. doi:10.1093/biostatistics/1.1.49.
- Dennis, M., P. Sandercock, J. Reid, C. Graham, J. Forbes, G. Murray, CLOTS (Clots in Legs Or sTockings after Stroke) Trials Collaboration. 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 (9891):516–524.
- Dennis, M., P. Sandercock, C. Graham, J. Forbes, J. Smith, and C. T. Collaboration. 2015. 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 Technology Assessment (Winchester, England)* 19 (76):1. doi:10.3310/hta19760.
- Ensor, H., R. J. Lee, C. Sudlow, and C. J. Weir. 2016. Statistical approaches for evaluating surrogate outcomes in clinical trials: A systematic review. *Journal of Biopharmaceutical Statistics* 26 (5):859–879. doi:10.1080/10543406.2015.1094811.
- Firth, D. 1993. Bias reduction of maximum likelihood estimates. *Biometrika* 80 (1):27–38. doi:10.1093/biomet/80.1.27.
- Frangakis, C. E., and D. B. Rubin. 2004. Principal stratification in causal inference. *Biometrics* 58 (1):21–29. doi:10.1111/j.0006-341X.2002.00021.x.
- Joe, H. 1989. Relative entropy measures of multivariate dependence. *Journal of the American Statistical Association* 84 (405):157–164. doi:10.1080/01621459.1989.10478751.
- Kent, J. T. 1983. Information gain and a general measure of correlation. *Biometrika* 70 (1):163–173. doi:10.1093/biomet/70.1.163.
- Kullback, S. 1997. *Information theory and statistics*. New York, NY: Dover publications.
- Linfoot, E. H. 1957. An informational measure of correlation. *Information and Control* 1 (1):85–89. doi:10.1016/S0019-9958(57)90116-X.
- Molenberghs, G., H. Geys, and M. Buyse. 2001. Evaluation of surrogate endpoints in randomized experiments with mixed discrete and continuous outcomes. *Statistics in Medicine* 20 (20):3023–3038. doi:10.1002/(ISSN)1097-0258.
- Molenberghs, G., M. Buyse, H. Geys, D. Renard, T. Burzykowski, and A. Alonso. 2002. Statistical challenges in the evaluation of surrogate endpoints in randomized trials. *Controlled Clinical Trials* 23 (6):607–625. doi:10.1016/S0197-2456(02)00236-2.
- Molenberghs, G., T. Burzykowski, A. Alonso, P. Assam, A. Tilahun, and M. Buyse. 2008. The meta-analytic framework for the evaluation of surrogate endpoints in clinical trials. *Journal of Statistical Planning and Inference* 138 (2):432–449. doi:10.1016/j.jspi.2007.06.005.
- Pryseley, A., A. Tilahun, A. Alonso, and G. Molenberghs. 2007. Information-theory based surrogate marker evaluation from several randomized clinical trials with continuous true and binary surrogate endpoints. *Clinical Trials* 4 (6):587–597. doi:10.1177/1740774507084979.
- Pryseley, A., A. Tilahun, A. Alonso, and G. Molenberghs. 2011. An information-theoretic approach to surrogate-marker evaluation with failure time endpoints. *Lifetime Data Analysis* 17 (2):195–214. doi:10.1007/s10985-010-9185-6.
- Renard, D., H. Geys, G. Molenberghs, T. Burzykowski, and M. Buyse. 2002. Validation of surrogate endpoints in multiple randomized clinical trials with discrete outcomes. *Biometrical Journal* 44 (8):921–935. doi:10.1002/bimj.200290004.
- Robins, J. M., and S. Greenland. 1992. Identifiability and exchangeability for direct and indirect effects. *Epidemiology* 3 (2):143–155. doi:10.1097/00001648-199203000-00013.
- Shannon, C.E. 1948. A mathematical theory of communication. In *a mathematical theory of communication*, vol. 27 doi:10.1002/bltj.1948.27.issue-4, 623–656. doi: 27 10.1002/bltj.1948.27.issue-4.
- Tilahun, A., A. Pryseley, A. Alonso, and G. Molenberghs. 2007. Flexible surrogate marker evaluation from several randomized clinical trials with continuous endpoints, using R and SAS. *Computational Statistics & Data Analysis* 51 (9):4152–4163. doi:10.1016/j.csda.2007.02.007.
- Tilahun, A., A. Pryseley, A. Alonso, and G. Molenberghs. 2008. Information theory-based surrogate marker evaluation from several randomized clinical trials with binary endpoints, using SAS. *Journal of Biopharmaceutical Statistics* 18 (2):326–341. doi:10.1080/10543400701697190.
- Van Swieten, J., P. Koudstaal, M. Visser, H. Schouten, and J. Van Gijn. 1988. Interobserver agreement for the assessment of handicap in stroke patients. *Stroke* 19 (5):604–607. doi:10.1161/01.STR.19.5.604.