Assessing the predictive value of a binary surrogate for a binary true endpoint based on the minimum probability of a prediction error

DOI: 10.1002/pst.1924
Handle: http://hdl.handle.net/1942/28995

Assessing the predictive value of a binary surrogate for a binary true endpoint, based on the minimum probability of a prediction error.

Paul Meyvisch, Ariel Alonso, Wim Van der Elst & Geert Molenberghs

Abstract

The individual causal association (ICA) has recently been introduced as a metric of surrogacy in a causal-inference framework. The ICA is defined on the unit interval and quantifies the association between the individual causal effect on the surrogate (ΔS) and true (ΔT) endpoint. In addition, the ICA offers a general assessment of the surrogate predictive value, taking value one when there is a deterministic relationship between ΔT and ΔS, and value zero when both causal effects are independent. However, when one moves away from the previous two extreme scenarios, the interpretation of the ICA becomes challenging. In the present work, we introduce the minimum probability of committing a prediction error when both endpoints are binary, i.e., the probability of erroneously predicting the value of ΔT using ΔS. This probability has a more straightforward interpretation but it is also shown that its magnitude is bounded above by a quantity that depends on the true endpoint. For this reason, the so-called reduction in prediction error (RPE) attributed to the surrogate is defined. The RPE can be more easily interpreted, it always lies in the unit interval, taking value 1 if prediction is perfect and 0 if ΔS conveys no information on ΔT. Furthermore, it has been demonstrated using simulations that the RPE is in strong agreement with the ICA. All analyses are illustrated using data from two clinical trials and a user-friendly R package *Surrogate* is provided to carry out the validation exercise.

Key words: Surrogate endpoint, Causal inference, Prediction Error, R package Surrogate

1 Introduction

Shortening the duration of clinical studies, reducing their economic costs and addressing their ethical issues are important motivations for the use of surrogate endpoints in clinical research [1]. Certainly, the extrapolation of the results obtained with a surrogate to the most clinically relevant outcome, the so-called true endpoint, cannot be made without risk. However, in some practical situations the use of surrogate endpoints may be the most reasonable strategy to move research forward [2]. Actually, as the surge of the AIDS and SARS epidemics in the past and the more recent expansion of Multi-Drug Resistant Tuberculosis in various parts of the world clearly showed, the potential of surrogate endpoints to speed up the approval of new therapeutic means may be of an immense value, in spite of their potential risk. In addition, the use of surrogate endpoints may be beneficial, not only in terms of cost or time, but they can also improve the accuracy in the estimation of target parametric functions such as the difference in success probabilities from different treatments, odds ratios and/or log risk ratios [3].

Over the last decades, several methodologies have been introduced for the evaluation of surrogate endpoints. Joffe and Greene grouped these methodologies into the so-called causal effects (CE) and causal association (CA) frameworks [4]. In these frameworks researchers have long tried to identify the

properties that a good surrogate should fulfill and several validation strategies have been introduced in the literature within the so-called causal-inference and meta-analytic paradigms[2,5,6,7]. In both paradigms attempts have been made to assess the capacity of the surrogate to predict the causal treatment effect on the true endpoint. For instance, when both endpoints are binary, Alonso *et al.* introduced an information-theoretic metric of surrogacy, the so-called individual causal association (ICA), to assess the surrogate predictive value[8]. To provide a more granular insight into the relationship between the individual causal treatment effect on the true endpoint ΔT and the individual causal treatment effect on the surrogate ΔS, Alonso *et al.* proposed the surrogate predictive function, which determines the most likely outcome of ΔT for any given value of ΔS[9]. Furthermore, these authors also introduced the so-called best prediction function, i.e., the function of ΔS that provides the best prediction for ΔT. In the present work, based on the best prediction function, the minimum probability of a prediction error is defined and an algorithm is proposed to handle the unidentifiability issues. In the rest of the manuscript, we will use the terms "minimum probability of a prediction error" and "probability of prediction error" (PPE) interchangeably. Furthermore, the so-called reduction in prediction error attributed to the surrogate (RPE) is defined and interpreted. It is shown that the RPE always lies in the unit interval, taking value 1 if prediction is perfect and 0 if ΔS conveys no information on ΔT. The RPE has a simple yet appealing interpretation in terms of the reduction in prediction error obtained from the surrogate, with respect to the prediction error obtained when only the information on the distribution of ΔT is used for prediction purposes. As both the RPE and the ICA are put forward as candidate measures to evaluate surrogate endpoints, it would be useful to investigate how both compare. An extensive simulation exercise is conducted which effectivily show that both measures are in strong agreement, i.e., higher values of the ICA correspond to higher values of the RPE and vice versa. In addition, both measures are fairly comparable in magnitude.

An additional objective of the current work is to present the analyses using the R package *Surrogate* as it is a tool that can help practitioners to conduct the complex analyses. All newly introduced concepts are built in the library that is freely available at CRAN. The remainder of this paper is organized as follows. In Section $2-4$, the theoretical model that underlies the causal-inference framework is detailed and the (reduction in) prediction error is introduced. The two case studies are introduced in Section 5 and the application of the concepts are provided in Section 6. In Section 7 some concluding remarks are provided.

2 Causal-inference model

The so-called Rubin's model for causal inference assumes that each patient has a four-dimensional vector of potential outcomes $\boldsymbol{Y} = (T_0, T_1, S_0, S_1)'$[10]. T_1, S_1, T_0 and S_0 are potential outcomes in that they represent the outcomes for the true (T) and surrogate (S) endpoint of an individual had he received the experimental treatment $(Z = 1)$ or control $(Z = 0)$, respectively. Each of the four variables is coded 1 (0) when a beneficial outcome is observed (not observed). We will temporarily restrict attention to the true endpoint, but similar arguments can be put forward for the surrogate endpoint as well.

The bivariate distribution of the vector of potential outcomes for the true endpoint $\boldsymbol{Y}_T = (T_0, T_1)'$ follows a multinomial distribution with parameters $\pi_{ij}^T = P(T_0 = i, T_1 = j)$ with $i, j = 0, 1$, and marginals $\pi_{i.}^T = \sum_j \pi_{ij}^T$, $\pi_{.j}^T = \sum_i \pi_{ij}^T$. Typically, only one of the two potential outcomes T_0 and T_1 can be observed and, consequently, the distribution of $\boldsymbol{Y}_T$ is often not identifiable[11] which implies that the association structure of the two potential outcomes cannot be inferred from the data. However, the marginal probabilities $\boldsymbol{\pi}_T = \left(\pi_{0.}^T, \pi_{1.}^T, \pi_{.0}^T, \pi_{.1}^T\right)'$ are identifiable under fairly gen-

Table 1: *Distribution of* $\boldsymbol{\Delta} = (\Delta T, \Delta S)'$.

		ΔS			
		-1	0	1	
ΔT	-1	π^{Δ}_{-1-1}	π^{Δ}_{-10}	π^{Δ}_{-11}	$\pi^{\Delta T}_{-1}$
	0	π^{Δ}_{0-1}	π^{Δ}_{00}	π^{Δ}_{01}	$\pi^{\Delta T}_{0}$
	1	π^{Δ}_{1-1}	π^{Δ}_{10}	π^{Δ}_{11}	$\pi^{\Delta T}_{1}$
		$\pi^{\Delta S}_{-1}$	$\pi^{\Delta S}_{0}$	$\pi^{\Delta S}_{1}$	1

eral conditions. Indeed, under SUTVA, $T = ZT_1 + (1-Z)T_0$ and if the treatment assignment is independent of the potential outcomes ($\boldsymbol{Y}_T \perp Z$), then $\pi^T_{1\cdot} = E(T|Z=0)$ with $\pi^T_{0\cdot} = 1 - \pi^T_{1\cdot}$ and $\pi^T_{\cdot 1} = E(T|Z=1)$ with $\pi^T_{\cdot 0} = 1 - \pi^T_{\cdot 1}$. SUTVA basically states that the potential outcomes of an individual are not affected by the treatments received by other individuals in the study and that the observed outcome under treatment Z equals the corresponding potential outcome T_Z. In addition, due to the random treatment allocation, the assumption of independence $\boldsymbol{Y}_T \perp Z$ can typically be guaranteed in randomized clinical trials.

The individual causal effect of the treatment on the true endpoint can be defined as $\Delta T = T_1 - T_0$; it follows a multinomial distribution parametrized by $\pi^{\Delta T}_i = P(\Delta T = i) = \sum_{pq} \pi^T_{pq}$ with $i = -1, 0, 1$ and the sum taken over all sub-indexes p, q satisfying $q - p = i$. Note that, like for $\boldsymbol{Y}_T$, the distribution of the individual causal treatment effect on the true endpoint ΔT is not identifiable from the data. However, it only requires making one untestable assumption about the association structure of the potential outcomes to identify it. Actually, it can easily be shown that assuming a specific value for π^T_{10}, is enough to fully identify the multinomial distribution of ΔT. Notice also that the range of π^T_{10} is constrained to the identifiable interval $[0, \min(\pi^T_{1\cdot}, \pi^T_{\cdot 0})]$.

Similarly, the potential outcomes $\boldsymbol{Y}_S = (S_0, S_1)'$ can be used to define the individual causal treatment effect on the surrogate endpoint ΔS and its distribution. Furthermore, the vector of individual causal treatment effects can be defined as $\boldsymbol{\Delta} = (\Delta T, \Delta S)'$; it follows the multinomial distribution given in Table 1 and it is the fundamental quantity used in the following sections to assess the surrogate PPE and RPE.

It has been argued that, if S is a good surrogate for T, then ΔS should convey a substantial amount of information about ΔT [8]. The mutual information between both individual causal treatment effects $I(\Delta T, \Delta S)$ quantifies precisely the average amount of uncertainty in ΔT, expected to be removed if the value of ΔS becomes known. Along these ideas, when both endpoints are binary, Alonso *et al.* proposed to quantify the ICA, i.e., the association between ΔT and ΔS, using the following transformation of the mutual information [8]

$$R^2_H(\Delta T, \Delta S) = \frac{I(\Delta T, \Delta S)}{\min\left[H(\Delta T), H(\Delta S)\right]},$$

where

$$I(\Delta T, \Delta S) = \sum_{i,j=-1}^{1} \pi_{ij}^{\Delta} \log\left(\frac{\pi_{ij}^{\Delta}}{\pi_i^{\Delta T}\pi_j^{\Delta S}}\right),$$

$$H(\Delta T) = -\sum_{i=-1}^{1} \pi_i^{\Delta T} \log(\pi_i^{\Delta T}),$$

$$H(\Delta S) = -\sum_{j=-1}^{1} \pi_j^{\Delta S} \log(\pi_j^{\Delta S}).$$

As previously stated, the first term $I(\Delta T, \Delta S)$ is the mutual information between both individual causal treatment effects and the other two expressions are the entropies of the individual causal treatment effects ΔT and ΔS. The concept of entropy lies at the center of information theory and quantifies the randomness or uncertainty associated with a random variable[12].

These authors showed that $R_H^2(\Delta T, \Delta S)$ is invariant under one to one transformations and that it always lies in the $[0, 1]$ interval, taking value zero when ΔT and ΔS are independent and value one when there is a nontrivial transformation ψ so that $P\left[\Delta T = \psi(\Delta S)\right] = 1$. Consequently, when $R_H^2(\Delta T, \Delta S) = 1$ there exists a deterministic relationship between both individual causal treatment effects, namely $\Delta T = \psi(\Delta S)$, and ΔS predicts ΔT without error. In addition, when $R_H^2(\Delta T, \Delta S) = 0$ both individual causal treatment effects are independent and no meaningful predictions are possible.

3 Minimum probability of a prediction error

In spite of being theoretically sound and intuitively appealing, the interpretation of the ICA may be challenging. For instance, it is difficult to define a cut-off point for the ICA, based on clinical considerations, that can help to identify good surrogate endpoints. In the present work we propose to quantify, for a given surrogate and the corresponding ICA, the PPE. Such a probability has an easy clinical interpretation and, therefore, it may be useful to identify good surrogates, where good can be defined in terms of the probability of making an incorrect assessment about ΔT based on ΔS. A low PPE is a necessary condition to qualify a surrogate and it should therefore be regarded as a measure that is complementary to the ICA. The PPE is derived from the best prediction function. Indeed, Alonso *et al.* defined the best prediction function associated with the distribution of $\mathbf{\Delta}$ as the function $\psi_b : \{-1, 0, 1\} \to \{-1, 0, 1\}$ satisfying $\psi_b = \arg\max_{\psi} P\left[\Delta T = \psi(\Delta S)\right]$[9]. These authors further proved that $\psi_b(j) = \arg\max_i P\left(\Delta T = i | \Delta S = j\right)$. If the argument function in the previous equation returns more than one value then any of them can be chosen arbitrarily to define $\psi_b(j)$; in such a case ψ_b will not be unique. Obviously, the best prediction function will also provide the minimum PPE. Actually, once the best prediction function ψ_b is obtained, the PPE is defined as:

$$P_e(\Delta T | \Delta S) = 1 - P\left[\Delta T = \psi_b(\Delta S)\right]. \tag{1}$$

If the information on the surrogate is completely ignored, then the best prediction for ΔT has to be based solely on the distribution of ΔT and it will take the form $\arg\max_i P\left[\Delta T = i\right]$. Obviously, in such scenario, the aforementioned prediction will have the smallest probability of a prediction error $P_e(\Delta T) = 1 - \max_i P\left[\Delta T = i\right]$. Ignoring the information on the surrogate is basically equivalent to using a prediction function ψ_m defined as $\psi_m(j) = \arg\max_i P\left[\Delta T = i\right]$ for all j. It follows from the definition of the best prediction function that

$$P\left[\Delta T = \psi_b(\Delta S)\right] \geq P\left[\Delta T = \psi_m(\Delta S)\right] = \max_i P\left[\Delta T = i\right], \tag{2}$$

and, therefore, $P_e(\Delta T|\Delta S) \leq P_e(\Delta T)$. As expected, it is very easy to show that the equality is reached if and only if ΔT and ΔS are independent. The previous inequality formalizes the intuitive idea that, if used correctly, additional information can only improve prediction. In practice, calculation of the reduction of the prediction error proceeds as follows. As the ICA is defined as a normalized version of the mutual information $I(\Delta T, \Delta S) = H(\Delta T) - H(\Delta T|\Delta S)$, it seems sensible to work with the reduction in the probability of a prediction error in a similar fashion, i.e., the RPE is defined as

$$RPE = \frac{P_e(\Delta T) - P_e(\Delta T|\Delta S)}{P_e(\Delta T)} = 1 - \frac{P_e(\Delta T|\Delta S)}{P_e(\Delta T)}, \tag{3}$$

The RPE quantifies how much the probability of a prediction error is reduced when using the surrogate as a predictor, with respect to the prediction based on the marginal distribution of ΔT only. From the previous developments it follows that the RPE always lies in the unit interval, taking value zero when ΔS conveys no information on ΔT and value one when both causal effects are deterministically related.

4 Identifiability issues

Causal inference models, based on potential outcomes, are conceptually attractive for the evaluation of surrogate endpoints. However, their use poses some practical challenges. Indeed, due to the so-called fundamental problem of causal inference, metrics of surrogacy developed from these models are not identifiable[11]. These identifiability problems are often tackled by defining identifiability conditions. For instance, to identify the estimands of interest, Gilbert and Hudgens, and Wolfson and Gilbert assumed that the surrogate endpoint under the control (S_0) was fixed and known[6,13]. Identifiability conditions are frequently combined with additional modeling assumptions in order to estimate the parameters of interest.

Although appealing, the use of identifiability conditions in this context has some conceptual and practical problems. In fact, often there is not enough substantive knowledge to assess the validity of the identifiability and/or modeling assumptions and, in general, they can be neither proven nor disproven based on the data. Therefore, the implementation of a sensitivity analysis, as proposed in Alonso *et al.*, seem to be more appropriate in this setting[8].

Alonso *et al.* approach the identifiability problem following a two-step Monte Carlo procedure, and based on the distribution of the vector of potential outcomes $\boldsymbol{Y}$. Actually, the parameter space of the distribution of $\boldsymbol{Y}$ is given by $\Gamma = \left\{\boldsymbol{\pi} \in [0,\ 1]^{16} :\ \mathbf{1}\boldsymbol{\pi} = 1\right\}$, where $\mathbf{1}$ is a vector of ones, $\boldsymbol{\pi} = (\pi_{ijpq})$, $\pi_{ijpq} = P(T_0 = i, T_1 = j, S_0 = p, S_1 = q)$ and $i, j, p, q = 0/1$. Due to the unidentifiability of $\boldsymbol{Y}$ the maximum likelihood estimator (MLE) of $\boldsymbol{\pi}$ is not unique, i.e., there is an entire region of the parameter space associated with the distribution of $\boldsymbol{Y}$, where the likelihood is maximized (Γ_D). In order to characterize Γ_D let us notice first that, as described in Li, Taylor and Elliott and in the work of Elliott, Li and Taylor, the data at hand impose some restrictions on π_{ijpq}[21,22]. In fact, the data allow identifying three probabilities $P(T = t, S = s|Z)$ within each treatment group and, thus, the 16 parameters characterizing the distribution of $\boldsymbol{Y}$ are subjected to 7 restrictions, implying that 9 are

allowed to vary freely and, hence, are not identifiable from the data. The set of restrictions on $\boldsymbol{\pi}$ can be written as:

$$\begin{aligned}
\pi_{1\cdot 1\cdot} &= P(T=1,S=1|Z=0), \quad \pi_{\cdot 1\cdot 1} = P(T=1,S=1|Z=1),\\
\pi_{1\cdot 0\cdot} &= P(T=1,S=0|Z=0), \quad \pi_{\cdot 1\cdot 0} = P(T=1,S=0|Z=1),\\
\pi_{0\cdot 1\cdot} &= P(T=0,S=1|Z=0), \quad \pi_{\cdot 0\cdot 1} = P(T=0,S=1|Z=1),\\
\pi_{\cdot\cdot\cdot\cdot} &= 1,
\end{aligned} \tag{4}$$

with the points in the sub-indexes indicating sums over those specific sub-indexes. Further, if one defines the vector $\boldsymbol{b}' = (1, \pi_{1\cdot 1\cdot}, \pi_{1\cdot 0\cdot}, \pi_{\cdot 1\cdot 1}, \pi_{\cdot 1\cdot 0}, \pi_{0\cdot 1\cdot}, \pi_{\cdot 0\cdot 1})$, then all the identified restrictions in (4) can be written as a system of linear equations $\mathbf{A}\boldsymbol{\pi} = \boldsymbol{b}$, with $\mathbf{A}$ a binary matrix. This hyperplane geometrically characterizes the subspace of Γ compatible with the data at hand, i.e., $\Gamma_D = \{\boldsymbol{\pi} \in \Gamma : \mathbf{A}\boldsymbol{\pi} = \boldsymbol{b}\}$. Essentially, the data at hand do impose some restrictions on $\boldsymbol{\pi}$ but do not fully determine it. The vector $\boldsymbol{b}$ contains all the estimable margins of $\boldsymbol{\pi}$ and in the original implementation of the algorithm these components were replaced by their MLE.

Furthermore, in a second step, the behavior of the parameters of interest like, for instance, the ICA is studied on Γ_D. Studying the behavior of a function on a region of an Euclidean space is a deterministic problem. However, using graphical or analytical techniques in this scenario is rather cumbersome due to the complex dependence of the ICA, PPE and RPE on $\boldsymbol{\pi}$ and the high dimensionality of the latter. Alonso *et al.* tackled these problems using a Monte Carlo approach. Monte Carlo methods are often used for obtaining numerical solutions to problems too complicated to solve analytically, like solving high-dimensional integrals, complex optimization problems or solving complex differential equations. Basically, in the second step, points are uniformly sampled on Γ_D and the estimads of interest are computed for all of them. Given that all points in Γ_D are equally compatible with the data, the use of a uniform sampling scheme is the most natural choice and it also guarantees that all regions on the hyperplane have the same probability of being covered by the sampling procedure.

As previously stated, in the original implementation of the algorithm, the sampling variability in the estimates of the marginal probabilities contained in $\boldsymbol{b}$ was not taken into account. Although this may only be a minor issue in large clinical trials, it may induce a non-negligible bias in small studies. Alonso *et al.* carried out a simulation study to evaluate this issue. They found that only when the sample size was rather small, i.e., $N = 50$ patients, certain degree of bias was observed. For instance, when estimating the ICA the relative bias was merely 3.5% and 1.3% for a sample size of $N = 100$ and $N = 300$, respectively[9]. If considered opportune, the sampling variability can be taken into account by uniformly sample the components of $\boldsymbol{b}$ from their corresponding confidence intervals at each run of the Monte Carlo algorithm[8,9]. This strategy is also implemented in the *Surrogate* package and in the Web appendix a new analysis of the case study is provided using this correction. We remit the interested reader to the original publications for a more detailed explanation of the procedure.

5 Case studies

In the present section the previous ideas will be used to assess the predictive value of two surrogate endpoints in ophthalmology and psychiatry. Both case studies have been previously analysed[8,9] to introduce the individual causal association and the surrogate predictive value respectively. We will re-analyse both case studies introducing the concepts of PPE and RPE, hereby providing more insights in the suitability of the proposed surrogate endpoints. We will additionally discuss a third case study in hepatitis B. It is also important to point out that a major practical problem frequently

encountered when validating surrogate endpoints, is the lack of user-friendly software packages to conduct the analysis. The R package *Surrogate* is freely available at `http://cran.r-project.org/web/packages/Surrogate/` and has been documented in [14]. The R package *Surrogate* additionally allows for the computation of the PPE and RPE. For conciseness, in the present section only a summary of the main results is given and no reference to the software is made. In the Supplementary Materials accompanying the paper a more detailed analysis of the case studies is provided and their implementation in R is discussed.

5.1 Collaborative Initial Glaucoma Treatment Study (CIGTS)

The Collaborative Initial Glaucoma Treatment Study (CIGTS) was a randomized clinical trial designed to evaluate the efficacy of surgery versus a conventional therapy in the treatment of patients suffering from glaucoma. A total of 228 patients were randomized to either surgery ($Z = 1$, 102 patients) or the conventional therapy ($Z = -1$, 126 patients). Both treatments were intended to bring intraocular pressure (IOP) down to less than 18 mm Hg. The surrogate endpoint was defined in terms of IOP at 12 months and the true endpoint at 96 months. S and T were equal to 1 if IOP was less than 18 mm Hg and to 0 otherwise[15]. See Table 2 for the cross-classification of surrogate and true endpoint for each treatment group. The data has been analysed by Alonso *et al.* who concluded that S is a poor surrogate for T based on low values of R^2_H [8]. A summary of the data is provided in top part of Table 2.

Table 2: Cross-tabulation of S versus T in the control (left) and experimental (right) treatment groups for both case studies.

CIGTS Case Study							
		Control				Experimental	
		T				T	
		0	1			0	1
S	0	36	32	S	0	15	9
	1	15	43		1	8	70
Psychiatric Case Study							
		T				T	
		0	1			0	1
S	0	105	12	S	0	94	7
	1	12	94		1	11	116

5.2 Psychiatric study

The data come from a clinical trial designed to compare the efficacy of risperidone (experimental group) and haloperidol (control group) in the treatment of schizophrenic patients. A total of $N = 454$ patients were treated for eight weeks and their condition was assessed using two psychiatric rating scales. Oftentimes in psychiatry, several rating scales are available to assess a patient's global condition. A useful and sufficiently sensitive assessment scale is the Positive and Negative Syndrome Scale (PANSS; [16]). PANSS consists of 30 items that provide an operationalized, drug-sensitive instrument, which is highly useful for both typological and dimensional assessment of schizophrenia. The Brief Psychiatric Rating Scale (BPRS; [17]) is a subscale of PANSS including only 18 items.

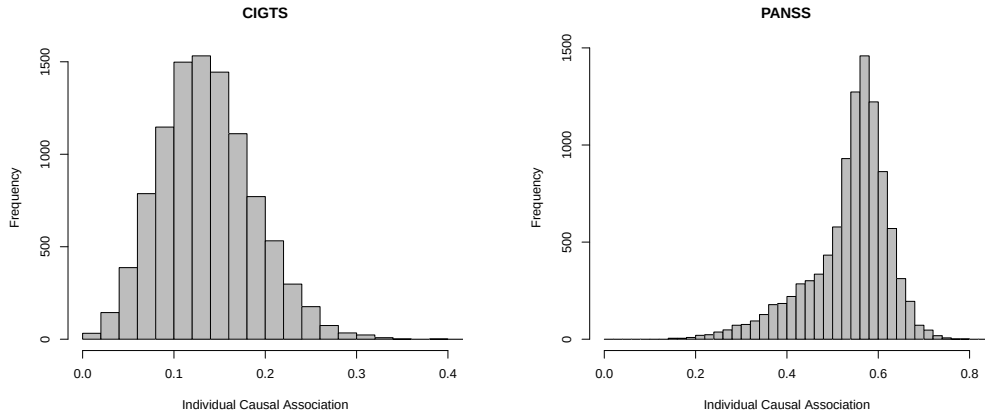

Figure 1: Individual Causal Association

The outcome of interest was the presence of a clinically relevant change in schizophrenic symptomatology as evaluated by the BPRS/PANSS scales. Clinically relevant change is defined as a reduction of 20% or more in the BPRS/PANSS scores, i.e, 20% reduction in post-treatment scores relative to baseline scores[18,19]. The BPRS is a scale which is more convenient to use than PANSS and it was investigated whether BPRS can be shown to be a reliable surrogate for PANSS. A summary of the data is provided in bottom part of Table 2.

5.3 Hepatitis B Surface Antigen (HBsAg)

In a recently concluded trial[20], comparing the combination of peginterferon $\alpha 2a$ (PEG) with Tenofovir Disoproxil Fumarate (TDF) versus PEG alone in patients with chronic hepatitis B, it was shown that combination therapy with PEG+TDF resulted in higher rates of Hepatitis B Surface Antigen (HBsAg) loss (8.6%) at the end of the trial (Week 72) compared to PEG alone (2.7%). What distinguishes this case study from the previous 2 is that the rates for T are very low. Unfortunately, the authors did not suggest a surrogate marker implying that the evaluation of the ICA, PPE and RPE can not be performed. The data are primarily used to illustrate the consequences of the low responder rates in regards to the expected magnitude of PPE for potential surrogate markers.

6 Results

The two-step Monte Carlo procedure introduced by Alonso *et al.* was applied to assess the estimands of interest[8,9]. Basically, a large number of vectors $\boldsymbol{\pi}$ were uniformly sampled on Γ_D, i.e., on the region of the parametric space of the distribution of $\boldsymbol{Y}$ compatible with the data at hand. From the set of obtained valid vectors $\boldsymbol{\pi}$, the distribution of $\boldsymbol{\Delta}$, given in Table 1, can be determined and, based on it, all the parameters of interest can be computed. Finally, frequency distributions for all the estimands can be obtained. These frequency distributions characterize the estimands across all scenarios compatible with the data and quantify the uncertainty emanating from the essential unidentifiably of the distribution of $\boldsymbol{Y}$.

Figure 1, left panel, shows the frequency distribution of the ICA values for the CIGTS. In general, the ICA seems to take rather small values on the entire Γ_D region. Indeed, as shown in Table 3,

50% of all the $\boldsymbol{\pi}$ vectors uniformly sampled on Γ_D led to ICA values smaller than 13% and for 95% of the sampled vectors the ICA never exceeded 23%. Therefore, looking at the previous results, a preliminary conclusion that can be drawn is that the IOP at 12 months is likely a poor surrogate for IOP at 96 months.

Li, Taylor and Elliott analyzed the CIGTS based on the so-called *associative* (AE) and *dissociative* (DE) effects[21]. Frangakis and Rubin introduced a *principal stratification* approach to evaluate surrogacy and suggested that the quality of a surrogate should be assessed based on the size of its *associative effect* relative to its *dissociative effect*[23]. The effect is associative if the causal treatment effect on T is reflected on the causal treatment effect on S, otherwise it is dissociative. A good surrogate is expected to have a large AE, indicating that the causal treatment effect on the surrogate is highly associated with the causal treatment effect on the true endpoint. Similarly, a good surrogate is expected to have a small DE, indicating that the causal treatment effect on the true endpoint is small when the causal treatment effect on the surrogate is zero[21,22]. Because AE and DE are constrained, Taylor, Wang and Thiébaut proposed to use instead the so-called associative (AP) and dissociative (DP) proportions, respectively[5].

Using log-linear models within a Bayesian framework, Li, Taylor and Elliott obtained results qualitative similar to our findings based on AP and DP[21]. However, it has been pointed out that AP and DP suffer from some conceptual problems. For instance, Alonso *et al.* considered the setting in which ΔT and ΔS are independent, i.e., the individual causal treatment effect on the surrogate conveys no information whatsoever on the individual causal treatment effect on the true endpoint[9]. Clearly, such a surrogate should not be considered valid. The results obtained from the ICA were conclusive, $R^2_H = 0$, i.e., knowing the individual causal treatment effect on the surrogate does not reduce our uncertainty about the individual causal treatment effect on the true endpoint at all. However, AP and DP could take any possible value in this setting depending on the value of $\pi_0^{\Delta S}$. Other problems were also detected in other scenarios. Based on these results Alonso *et al.* concluded that, at least in some scenarios, the ICA offers a more coherent assessment of surrogacy than AP and DP[9].

The results for the PANSS data are more uncertain. Indeed, as shown in the right panel of Figure 1, the ICA can take both small and moderately high values on Γ_D (see Table 3). In fact, while for 5% of the sampled vectors the ICA was smaller than 36%, for other 5% it was larger than 65% and the frequency distribution has a median value of 56%. It is in this kind of "middle range" scenarios that the interpretation of the ICA becomes difficult and the proposed PPE and RPE can be of great value. Actually, it is difficult to determine, in clinical terms, how large an ICA values need to be to establish surrogacy. In general, most data analysts would agree that values larger than 90% or smaller than 20% offer evidence of good and poor surrogacy respectively, but "middle range" values are certainly hard to interpret.

6.1 Distribution of prediction functions

Figure 2 shows the frequencies at which different functions were selected as the best prediction function. Basically, for every valid sampled $\boldsymbol{\pi}$ vector on Γ_D, the distribution of $\Delta = (\Delta T, \Delta S)$ was first obtained, i.e, the cell probabilities in in Table 1 were determined. Subsequently, for every column, i.e., for every value of ΔS, the row with the largest probability was selected. This way, 3 cells were selected, one in each column. The selection of these 3 cells represents the best prediction function for the sampled $\boldsymbol{\pi}$ and the sum of the 6 cells that are not selected gives the probability of making a prediction error PPE. There exist a total of 27 possible prediction functions $\psi : \{-1, 0, 1\} \rightarrow \{-1, 0, 1\}$ which for ease of notation are represented by a triplet (a, b, c) with $\psi(-1) = a$, $\psi(0) = b$

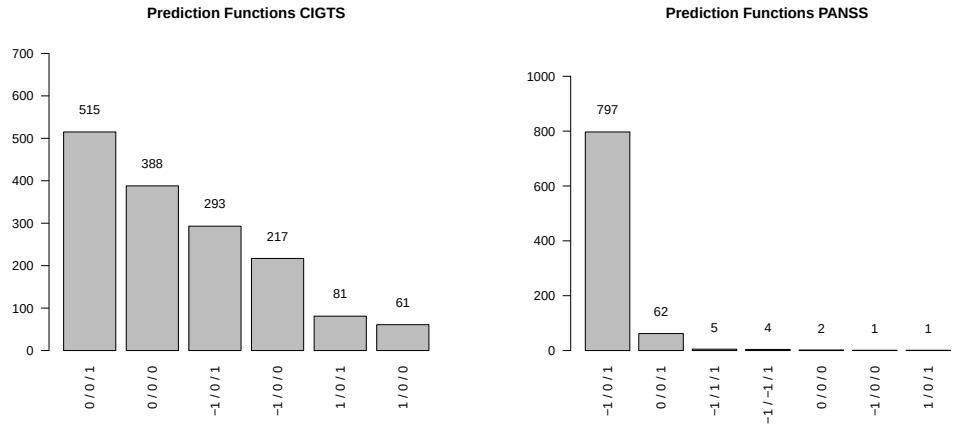

Figure 2: Most frequently selected prediction functions

and $\psi(1) = c$. For example, the triplet $(-1, 0, 1)$ represents the identity function, i.e., projecting every value of ΔS onto the same value of ΔT. Similarly, the triplet $(0, 0, 0)$ represents the function that projects every value of ΔS onto $\Delta T = 0$, etc. From each valid vector of $\boldsymbol{\pi}$ that is sampled, the best prediction function represented by the corresponding triplet (a, b, c) is obtained. Sampling a large number of $\boldsymbol{\pi}$ vectors yields a distribution of triplets which reveals an insightful pattern. For instance, a very clear picture is observed in the PANSS study. Here, the identity function $(-1, 0, 1)$ is by far the most frequently selected from which it follows that the best prediction function and the identity function largely overlap. Given that BPRS is essentially a subscale of PANSS, this result is intuitively plausible.

The results for CIGTS trial are substantially different. In this case even the most frequently selected prediction function $(0, 0, 1)$ is selected in less than 50% of the times. In addition, the function $(0, 0, 0)$ is the second most frequently selected which is yet another indication that S may be a poor surrogate for T. Indeed, for all samples $\boldsymbol{\pi}$ that are obtained for which $(0, 0, 0)$ is the best prediction function, ΔS does not provide information that improves the prediction of ΔT that one would obtain based only on the distribution of the latter, i.e., ΔS can be completely ignored. Notice that these results are completely in line with the conclusions obtained from the analysis of the ICA values.

6.2 The PPE and RPE

The distribution of the PPE and the RPE are graphically presented in Figure 3 and Figure 4 respectively. Summary statistics of the distribution of PPE and RPE are provided in Table 3. As the top part of the table illustrates, in the CIGTS for 95% of the sampled vectors the PPE exceeds 28% and for 50% of them it is larger than 37%. This is underscored by the distribution of the RPE which has a median value of 8.8%, indicating that there is only marginal gain in predictivity when using the information of ΔS to predict ΔT. It is also observed that for a substantial part of the distribution, the RPE equals zero, which is consistent with the earlier finding that the prediction function $(0, 0, 0)$, which is independent from ΔS, was often selected as the best prediction function. The previous results, together with the low ICA values already found, strongly suggest that the IOP at 12 months is indeed a poor surrogate for IOP at 96 months.

The analysis of the PANSS data is even more interesting. Indeed, as the bottom part of Table 3 clearly shows, for 95% of the sampled vectors the PPE does not exceed 16% and for 50% of them it

is actually smaller than 13.5%. In addition, the high values for RPE indicate that there is a subtantial reduction in PPE when ΔT is predicted using the information conveyed by ΔS. Notice that, for 95% of the sampled vectors, using ΔS to predict ΔT reduced the PPE in more than 32%. Whether or not BPRS qualifies as a suitable surrogate for PANSS warrants a clinical discussion but it does put the magnitude of the ICA in a much better perspective. Furthermore, parameters like the PPE and RPE are much easier to understand and, therefore, it would be much easier for clinicians to define cut off points for them in order to establish surrogacy. // Another useful analysis consists of the assessment of association between ICA and RPE. We refer to the Supplementary Materials for a more detailed analysis. It is reassuring that the association in the PANSS case study is quite high as the Pearson correlation coefficient between ICA and RPE is 93%. A Pearson correlation coefficient of 61% has been observed in the CIGTS case study.

Finally, as there is no proposed surrogate for the HBsAg study, densities for ICA, PPE and RPE can not be produced. However, it has been demonstrated in (2) that the PPE is always bounded above by the probability of the prediction error based on the marginal distribution of ΔT only, which we have denoted as $P_e(\Delta T)$. It is a striking result that any surrogate that is to be proposed, including surrogates with a very low ICA, will yield a median PPE of less than 8.6%. However, this is a feature of the trial results, as very low response rates have been observed, and is independent from the choice of surrogate. This should not come as a surprise as every clinician treating hepatitis B patients is well aware that the current therapies fall short in attaining sustained HBsAg loss. It is therefore a relatively safe bet that an individual patient will fail treatment, irrespective which of the two therapies are assigned. Translating this to the causal inference setting, $\Delta T = 0$ is always the best prediction which is associated with a high probability. Conversely, the probability of making a prediction error using information on the clinical trial results for T only, is low and equals $1 - P(\Delta T = 0)$ for which the values are provided in Table 3. As earlier stated, the PPE for any proposed surrogate will be bounded above by these values.

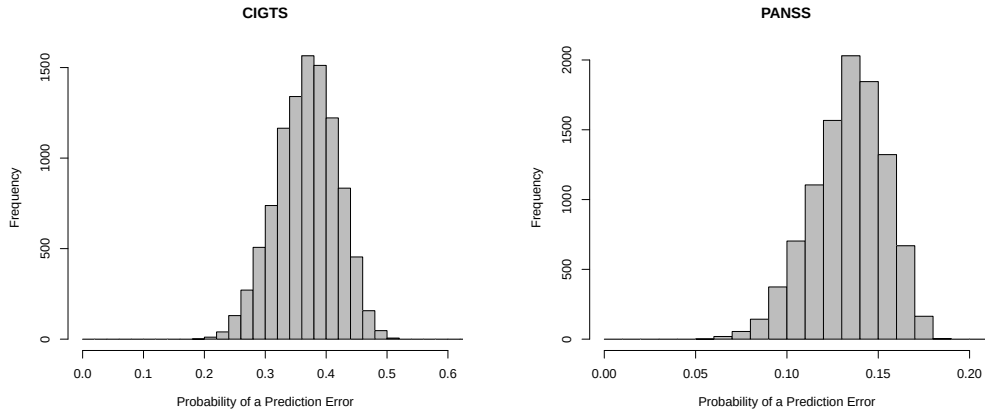

Figure 3: Probability of a Prediction Error

6.3 Simulations

A simulation exercise was conducted to explore the relationship between the ICA and RPE. To this end, a large set of trials are simulated under various scenarios and the ICA and RPE values that emerged from these simulated trials have been systematically compared. We highlight the conclusions of this exercise in this section and refer the interested reader to the Supplementary Materials. More

Table 3: Distribution of PPE and RPE

CIGTS	5%	10%	20%	50%	80%	90%	95%
ICA	0.057	0.073	0.092	0.133	0.178	0.205	0.227
PPE	0.282	0.301	0.326	0.370	0.410	0.431	0.446
RPE	0	0	0	0.088	0.194	0.246	0.288
PANSS							
ICA	0.357	0.411	0.479	0.556	0.602	0.626	0.648
PPE	0.098	0.107	0.117	0.135	0.151	0.159	0.164
RPE	0.319	0.418	0.548	0.725	0.777	0.798	0.815
HBsAg							
PPE	< 0.062	< 0.064	< 0.070	< 0.086	< 0.102	< 0.108	< 0.110

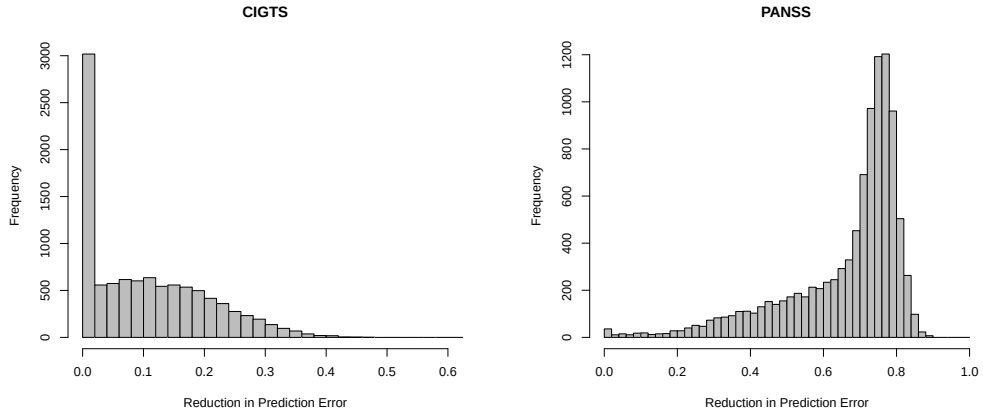

Figure 4: Reduction of the Prediction Error

specifically, we focused on whether the densities of ICA and RPE are strongly correlated, i.e., whether high (low) values of ICA in Figure 1 effectively associate with high (low) values of the RPE in Figure 4. Another objective of this exercise consisted in comparing central tendency of both densities. The simulation results have indeed reliably demonstrated that there is strong relationship between the ICA and RPE in the sense that the mean values consistently move in the same direction, i.e., surrogates having higher mean ICAs will naturally have higher mean RPEs and vice versa. In addition it is fair to conclude that both mean values are numerically close to one another. As far as the pairwise correlation between the ICA and RPE values for a given trial is concerned, a more complex pattern is observed which is in line with the findings on the PANSS and CIGTS studies. The results have clearly indicated that there is strong correlation in most settings with some exceptions. First, the correlation can be low when the width of the ICA and RPE densities are narrow which is of limited concern. Indeed, when this occurs, the ICA and RPE can reliably be represented by its mean values which we have shown to be consistent. Another reason impacting the correlation is when the ICA and RPE densities are low in magnitude. This is related to the intrinsic nature of the PPE and RPE. Similar to what has been depicted in Figure 2, there is a wider range of prediction functions selected when the surrogate is poor. In the extreme case, when prediction function $(0, 0, 0)$ is frequently selected, the RPE is consistently equal to zero. The more discrete nature of RPE negatively impacts the correlation with the ICA. In light of above findings we can conclude that the Pearson correlation for the PANSS study was high (93%) due to the relatively large width of the ICA and RPE densities in addition

to the high magnitude of the ICA density. Conversely, the Pearson correlation for the CIGTS study was lower (61%) due to the more narrow width of the ICA and RPE densities along with the lower magnitude of the ICA density.

7 Conclusions

The ICA has recently been proposed as a metric of surrogacy in a causal-inference framework. It is defined on the unit scale and takes the value 1 when there exists a deterministic relationship between ΔS and ΔT and value zero when both causal effects are independent. However, in practical settings, the ICA will take values somewhere on the unit interval and it is challenging to define thresholds for the ICA in the absence of a clear clinical interpretation. This calls for development of additional measures such as the PPE that express surrogacy in terms of the probability of making a prediction error, which has a more straightforward interpretation. Actually, clinicians may often be able to define the risk they are willing to take, when using a surrogate, in terms of the probability prediction error. For instance, they may determine that a surrogate that leads to erroneous predictions of the individual causal treatment effect on the true endpoint in less than 20% of the cases, is acceptable in certain medical contexts. Analogously, they may determine that a surrogate that leads to erroneous predictions of the individual causal treatment effect on the true endpoint in more than 30% of the cases is not acceptable. From this perspective, the PPE is complementary to the ICA as it exactly equals the probability of making a prediction error on ΔT after accounting for the information that ΔS has on ΔT. One would therefore expect that high values of the ICA naturally correspond to low values of the PPE, as was seen in the PANSS case study. However, the HBsAg case study has revealed that caution has to be exercised in interpreting the PPE as low values may be the result of characteristics of the true rather than the surrogate endpoint. The fact that PPE is bounded above by a quantity that depends on the true endpoint and the treatment under consideration, clearly hinders its interpretation. Therefore, the RPE may often be more useful for the task at hand than the PPE. Indeed, the RPE quantifies how much the probability of a prediction error is reduced when using the surrogate as a predictor, with respect to the prediction based on the marginal distribution of ΔT only. The RPE is constructed in a similar fashion as the ICA but uses the marginal and conditional probability of a prediction error on ΔT instead of the marginal and conditional entropy of ΔT as primary building blocks. It shares with the ICA the convenient property that it takes value zero when ΔS conveys no information on ΔT and value one when both causal effects are deterministically related. Moving away from these extreme scenarios simulations have demonstrated that the RPE and the ICA will behave approximately similarly while the RPE has a more straightforward interpretation. Interestingly, as the field of Hepatitis B is moving quickly with new mechanisms of action that are being tested in Phase 1 and 2 trials, there are high hopes that the rates of therapeutic success will substantially increase in the next decade. This will inevitably have an impact on the relationship between RPE and PPE as the prediction based on the marginal distribution of ΔT will be associated with more uncertainty. As an example, if we were to impose a PPE of 5% with a $P_e(\Delta T) = 10\%$, it follows that RPE=50%. Similarly, with $P_e(\Delta T) = 50\%$, the RPE has to increase to 90%. This corresponds to a surrogate which has a near-deterministic relationship with the true endpoint. It can be concluded from this simple example that any requirement for a PPE and RPE should be based on both clinical and statistical arguments.

References

[1] Alonso A, Molenberghs G. (2008). Surrogate endpoints: Hopes and perils. *Pharmacoeconomics and Outcomes Research* ; **8**: 255–259. DOI:10.1586/14737167.8.3.255

[2] Burzykowski, T., Molenberghs, G., and Buyse, M. (2005). *The Evaluation of Surrogate Endpoints.* New York: Springer-Verlag.

[3] Banerjee, B. and Biswas, A. (2015) Linear increment in efficiency with the inclusion of surrogate endpoint. *Statistics and Probability Letters*, **96**, 102–108

[4] Joffe M.M. and Greene T. (2009). Related Causal Frameworks for Surrogate Outcomes. *Biometrics 65, 2*, 530–538.

[5] Taylor, J.M.G., Wang, Y., and Thiébaut, R. (2005). Counterfactual links to the proportion of treatment effect explained by a surrogate marker. *Biometrics*, **61**, 1102–1111.

[6] Gilbert, P.B. and Hudgens, M.G. (2008). Evaluating candidate principal surrogate endpoints. *Biometrics* **64**, 1146–1154.

[7] VanderWeele, T. (2013). Surrogate measures and consistent surrogates. *Biometrics*, **69**, 561–565.

[8] Alonso A, Van der Elst W, Molenberghs G, Buyse M and Burzykowski T. (2016). An information-theoretic approach for the evaluation of surrogate endpoints based on causal inference. *Biometrics*, **72, 3**, 669–677.

[9] Alonso A, Van der Elst W and Meyvisch P (2016). Assessing a surrogate predictive value: A causal inference approach. *Statistics in Medicine* (DOI: 10.1002/sim/.7194).

[10] Rubin, D. B. (1980). Randomization analysis of experimental-data the Fisher randomization test – comment. *Journal of the American Statistical Association* **75**, 591–593.

[11] Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association* **81**, 945–960.

[12] Joe H. (1989). Relative entropy measures of multivariate dependence. *Journal of the American Statistical Association* **84**, 157–164.

[13] Wolfson J and Gilbert P. (2010) Statistical identifiability and the surrogate end–point problem, with application to vaccine trials. *Biometrics* **66(4)**, 1153–1161.

[14] Alonso, A., Theophile Bigirumurame, Tomasz Burzykowski, Marc Buyse, Geert Molenberghs, Leacky Muchene, Nolen Joy Perualila, Ziv Shkedy, Wim Van der Elst. (2017) *Applied Surrogate Endpoint Evaluation Methods with SAS and R*. Chapman and Hall/CRC.

[15] Musch, D. C., Lichter, P. R., Guire, K. E., Standari, C. L., and CIGTS Investigators (1999). The collaborative initial glaucoma treatment study: Study design, methods, and baseline characteristics of enrolled patients. *Ophthalmology, 43*, 137-160.

[16] Singh M, Kay S. (1975). A comparative study of haloperidol and chlorpromazine in terms of clinical effects and therapeutic reversal with benztropine in schizophrenia. Theorectical implications for potency differences among neuroleptics. *Psychopharmacologia* **43**, 103–113.

[17] Overall J, Gorham D. (1962). The Brief Psychiatric Rating Scale. *Psychological Reports* **10**: 799–812.

[18] Kane J, Honigfeld G, Singer J, Meltzer H. (1988). Clozapine for the treatment-resistant schizophrenic. A double-blind comparison with chlorpromazine. *Archives of General Psychiatry* **45**, 789–796.

[19] Leucht S, Kane JM, Kissling W, Hamann J, Etschel E, Engel R (2005). Clinical implications of the Brief Psychiatric Rating Scale Scores. *British Journal of Psychiatry* 2005; **187**: 366–371.

[20] Marcellin et al. (2016). Combination of Tenofovir Disoproxil Fumarate and Peginterferon $\alpha 2a$ Increases Loss of hepatitis B Surface Antigen in Patients with Chronic Hepatitis B. *Gastroenterology* 150:133-144.

[21] Li Y, Taylor JMG, Elliott MR. A Bayesian approach to surrogacy assessment using principal stratification in clinical trials. *Biometrics* 2010; **58**: 21–29.

[22] Elliott MR, Li Y, Taylor JMG. Accommodating missingness when assessing surrogacy via principal stratification. *Clinical Trials* 2013; **10**: 363–377.

[23] Frangakis C.E. and Rubin D.B. (2002). Principal stratification in causal inference. *Biometrics* **58**, 21–29.