A reflection on the possibility of finding a good surrogate

Peer-reviewed author version

ALONSO ABAD, Ariel; MEYVISCH, Paul; VAN DER ELST, Wim; MOLENBERGHS, Geert & VERBEKE, Geert (2019) A reflection on the possibility of finding a good surrogate. In: JOURNAL OF BIOPHARMACEUTICAL STATISTICS, 29(3), p. 468-477.

DOI: 10.1080/10543406.2018.1559854
Handle: http://hdl.handle.net/1942/30132

A reflection on the possibility of finding a good surrogate

Ariel Alonso[*,1] **Paul Meyvisch**[2]

Wim Van der Elst[2] **Geert Molenberghs**[3,1]

Geert Verbeke[1,3]

[1], I-BioStat, KU Leuven, Kapucijnenvoer 35, B3000 Leuven, Belgium.

[2], Janssen Pharmaceutical, Companies of Johnson & Johnson, Belgium.

[3], I-BioStat, Universiteit Hasselt, Agoralaan, B3590 Diepenbeek, Belgium.

[*] *E-mail:* ariel.alonsoabad@kuleuven.be

Abstract

Surrogate endpoints need to be statistically evaluated before they can be used as substitutes of true endpoints in clinical studies. However, even though several evaluation methods have been introduced over the last decades, the identification of good surrogate endpoints remains practically and conceptually challenging. In the present work, the question regarding the existence of a good surrogate is addressed using information-theoretic concepts, within a causal inference framework. The methodology can help practitioners to assess, given a clinically relevant true endpoint and a treatment of interest, the chances of finding a good surrogate endpoint in the first place. The methodology focuses on binary outcomes and is illustrated using data from the Initial Glaucoma Treatment Study. Furthermore, a newly developed and user friendly R package *Surrogate* is provided to carry out the necessary calculations.

Keywords: Causal inference, Fano's inequality, Surrogate endpoints.

1 Introduction

Surrogate endpoints like cholesterol, blood sugar levels, and blood pressure have enabled pharmaceutical companies and health professionals to carry out faster and more efficient clinical studies, for evaluating life-saving and health-promoting interventions. They have also improved our understanding of some disease processes and helped public health authorities to identify and track health concerns (Michel and Ball, 2010).

However, the use of surrogate endpoints has been controversial as well. For instance, long-term hormone replacement therapy significantly lowered "bad" cholesterol and raised "good" cholesterol in women but, at the same time, it increased their chances of heart attacks and strokes (Writing Group for the Women's Health Initiative, 2002). These unfortunate events made clear that surrogate endpoints need to be evaluated before they can be used as substitutes of true endpoints in clinical studies and led to the development of several evaluation strategies within the so-called causal-inference and meta-analytic paradigms (Joffe and Greene, 2009; Alonso *et al.*, 2017).

In spite of important methodological advances, the identification of good surrogate endpoints remains extremely challenging (Buyse *et al.*, 2000, 2010). In fact, as practice has shown, the evaluation of surrogate endpoints is often a strenuous process, with respect to both the initial demonstration of a relationship between a putative surrogate and the clinical endpoint, and its subsequent statistical validation (Buyse *et al.*, 2010). Therefore, addressing the existence question should be an important first step before embarking on the search for a good surrogate marker and, obviously, the very meaning of "good" needs to be rigorously defined. Essentially, one would like, given a clinically relevant true endpoint and a treatment of interest, to assess the plausibility of finding a good surrogate endpoint in the first place. To our

knowledge, little has been done to address this important issue. Alonso *et al.* (2015) studied the existence problem in the setting in which both endpoints are normally distributed, but many other relevant scenarios have not been investigated yet.

Methodologies have been developed for the evaluation of a binary outcome as a putative surrogate for a binary true endpoint (Gilbert and Hudgens, 2008; Li, Taylor and Elliott, 2010; Elliott, Li and Taylor, 2013). Recently, Alonso *et al.* (2016a) introduced an information-theoretic metric of surrogacy in the binary-binary setting, the so-called individual casual association (ICA), and addressed the identifiability issues using a two-step Monte Carlo procedure. However, the plausibility of finding a good surrogate in this important scenario has not been studied yet. In the present work, we address this important problem using information-theoretic concepts, more specifically, the so-called Fano's inequality.

In Section 2, a causal-inference model is introduced. In Section 3, an information-theoretic framework is presented to assess the likelihood of finding a good surrogate endpoint. The methodology presented in Section 3 is applied in Section 4 to analyze the likelihood of finding a valid surrogate endpoint in the context of glaucoma research. Finally, some concluding remarks are given in Section 5.

2 Causal-inference model

We will consider the setting in which both the true (T) and surrogate (S) endpoints are binary variables coded as 1 (0) when a beneficial outcome is observed (not observed) and only two treatments are under evaluation ($Z = 0/1$). In addition, the standard stable unit treatment value assumption (SUTVA) will also be made (Rubin , 1980).

Based on the so-called Rubin's model for causal inference it will be assumed that there exists, for each patient, a four dimensional vector of potential outcomes $\boldsymbol{Y} = (T_0, T_1, S_0, S_1)'$. The components T_1, S_1, T_0 and S_0 represent the outcomes for the true and surrogate endpoint of an individual had he received the treatment or control, respectively. In the following, the discussion will be temporarily restricted 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 $\boldsymbol{Y}_T = (T_0, T_1)'$ is parametrized by $\pi_{ij}^T = P(T_0 = i, T_1 = j)$ with $i, j = 0, 1$, and this parametrization leads to the marginals $\pi_{i\cdot}^T = \sum_j \pi_{ij}^T$, $\pi_{\cdot j}^T = \sum_i \pi_{ij}^T$. In practice, only one of the two potential outcomes T_0 and T_1 can be observed and the distribution of $\boldsymbol{Y}_T$ is therefore not identifiable (Holland, 1986). More specifically, the odds ratio $\theta_T = \pi_{00}^T \pi_{11}^T / \pi_{10}^T \pi_{01}^T$ quantifying the association between the two potential outcomes cannot be inferred from the data. Unlike θ_T, the marginal probabilities $\boldsymbol{\pi}_T = \left(\pi_{0\cdot}^T, \pi_{1\cdot}^T, \pi_{\cdot 0}^T, \pi_{\cdot 1}^T\right)'$ are identifiable under fairly general conditions. In fact, under SUTVA, $T = ZT_1 + (1 - Z)T_0$ and if the treatment assignment mechanism is independent of the potential outcomes ($\boldsymbol{Y}_T \perp Z$), then $\pi_{1\cdot}^T = E\left(T|Z = 0\right)$ with $\pi_{0\cdot}^T = 1 - \pi_{1\cdot}^T$ and $\pi_{\cdot 1}^T = E\left(T|Z = 1\right)$ with $\pi_{\cdot 0}^T = 1 - \pi_{\cdot 1}^T$. Importantly, due to the random treatment allocation, the aforementioned assumption of independence $\boldsymbol{Y}_T \perp Z$ can often be guaranteed in randomized clinical trials. In addition, once the odds ratio θ_T has been given a value, the full bivariate distribution of $\boldsymbol{Y}_T$ can be recovered using $\boldsymbol{\pi}_T$ (Plackett, 1965).

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 with parameters $\pi_i^{\Delta T} = P(\Delta T = i) = \sum_{pq} \pi_{pq}^T$, $i = -1, 0, 1$, where the sum is taken over all sub-indexes p, q satisfying $q - p = i$. Although the distribution of ΔT is not generally identifiable

from the data, once θ_T is fixed, it becomes fully identifiable.

The potential outcomes $\boldsymbol{Y}_S = (S_0, S_1)'$ can be similarly used to define the individual causal treatment effect on the surrogate ΔS and its distribution. The vector of individual causal treatment effects $\boldsymbol{\Delta} = (\Delta T, \Delta S)'$, which follows the multinomial distribution given in Table 1, is one of the fundamental concepts needed to assess the likelihood of finding a good surrogate endpoint.

3 An information-theoretic framework

It has been argued that understanding the association between the causal treatment effects on the true and surrogate endpoint is critical to understanding the value of a surrogate from a clinical perspective (Elliott, Li and Taylor, 2013). Along these lines, Alonso *et al.* (2015) proposed to assess surrogacy using the so-called individual causal association (ICA). When both endpoints are continuous and normally distributed, these authors quantified the ICA using the Pearson correlation coefficient $\rho_\Delta = \text{corr}\,(\Delta T, \Delta S)$. However, when one moves way from the realm of normality, assessing the association between the individual causal treatment effects, in an intuitively appealing way, becomes a more challenging task. Alonso *et al.* (2016a) used information-theoretic concepts to quantify the ICA when both endpoints are binary outcomes. Although the technical details need to be worked out for every outcome type combination, i.e., binary-binary, continuous-binary, among others, information theory offers a conceptual framework for the evaluation of surrogate endpoints across all these scenarios.

3.1 Information theory: Some important concepts

Information theory deals with the study of problems concerning complex systems, and has been applied in a variety of fields such as modern communication theory. In spirit and concepts, information theory has its mathematical roots connected to thermodynamics and statistical mechanics. The concept of entropy, introduced by Shannon (1948), quantifies the "epistemic" uncertainty or lack of knowledge implied by a distribution. Information and entropy are opposite concepts, i.e., to gain information is to lose uncertainty by the same amount and, hence, their formal definitions differ only in sign. For a discrete random variable Y with finite support $\{y_1, y_2, \ldots, y_m\}$ and probability function $P(Y = y_i) = \pi_i$, entropy is defined as

$$H(Y) = -E_Y[\log P(Y)] = -\sum_i \pi_i \log \pi_i.$$

The joint and conditional entropies can be defined in a similarly way as $H(X, Y) = -E_{X,Y}[\log P(X, Y)]$ and $H(Y|X) = -E_X[E_Y(\log P(Y|X))]$, with $P(x, y)$ and $P(y|x)$ denoting the joint and conditional probability functions, respectively. Entropy is always non-negative and satisfies $H(Y|X) \leq H(Y)$ for any pair of random variables (X, Y), with equality holding under independence. The foregoing inequality essentially states that, as an average, uncertainty on Y can only decrease if additional information (X) becomes available. Furthermore, entropy is invariant under a bijective transformation.

The mutual information $I(X, Y)$ quantifies the amount of uncertainty in Y, expected to be removed if the value of X were known, and it is defined as $I(X, Y) = H(Y) - H(Y|X)$. Mutual information is always non-negative, zero if and only if X and Y are independent, symmetric, invariant under bijective transformations of X and Y, and $I(X, X) = H(X)$. It follows from the definitions of entropy and mutual

information that

$$I(X,Y) = H(X) + H(Y) - H(X,Y) = \sum_{x,y} P(x,y) \log\left(\frac{P(x,y)}{P(x)\,P(y)}\right),$$

where $P(x,y)$, $P(x)$, and $P(y)$ denote the joint and marginal probability functions of X and Y, respectively.

Fano's inequality is an important result in information theory that relates the probability of error in predicting a random variable Y, based on another random variable X, to its conditional entropy $H(Y|X)$. If $\tilde{Y} = f(X)$ is the prediction of Y and $P_e = P(Y \neq \tilde{Y})$ denotes the probability of a prediction error, then Fano's inequality states that

$$H(X|Y) \leq H(P_e) + P_e \log(|\mathcal{X}| - 1)$$

where $\mathcal{X}$ denotes the support of X, $|\mathcal{X}|$ is its cardinality and

$$H(P_e) = -P_e \log P_e - (1 - P_e) \log(1 - P_e)$$

In the following sections these information-theoretic concepts will be applied to answer important questions regarding the validity of a putative surrogate and the likelihood of finding a valid surrogate endpoint.

3.2 Individual causal association: A review

Based on the previous concepts, Alonso *et al.* (2016a) introduced the following information-theoretic measure of association to assess the ICA when both endpoints

are binary outcomes

$$R_H^2(\Delta T, \Delta S) = \frac{I(\Delta T, \Delta S)}{\min\left[H(\Delta T), H(\Delta S)\right]}, \tag{1}$$

where the term in the numerator is given by

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

and the denominator in (1) equals the minimum of the entropies of the individual causal treatment effects $H(\Delta T) = \sum_{i=-1}^{1} \pi_i^{\Delta T} \log(\pi_i^{\Delta T})$ and $H(\Delta S) = \sum_{j=-1}^{1} \pi_j^{\Delta S} \log(\pi_j^{\Delta S})$. The mutual information between both individual causal treatment effects quantifies the amount of uncertainty in ΔT expected to be removed if the value of ΔS becomes known and, hence, it seems sensible to assess surrogacy based on this information-theoretic measure. Actually, using some theoretical arguments, Alonso *et al.* (2016a) showed that, at least in some scenarios, the ICA may offer a more coherent assessment of surrogacy than other previously introduced metrics. Furthermore, it can be shown that, under normality, the Pearson correlation coefficient ρ_{Δ} is a rescaled version of the mutual information.

The ICA, as given in (1), can also be interpreted as a measure of prediction accuracy, i.e., a measure of how accurately one can predict the causal treatment effect on the true endpoint for a given individual, using his causal treatment effect on the surrogate. Indeed, Alonso *et al.* (2016a) showed that $R_H^2(\Delta T, \Delta S)$ is invariant under one-to-one transformations and that it always lies in the unit 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 $R^2_H(\Delta T, \Delta S) = 0$ both individual causal treatment effects are independent and no meaningful predictions are possible. In the following section, this relationship between the ICA and prediction will be used to develop an approach to evaluate the existence problem, i.e., to assess the likelihood of finding a good surrogate endpoint in the first place.

3.3 Is there a good surrogate?

As stated in section 1, before addressing the existence question, one first needs to reflect on the very meaning of the term "good" surrogate endpoint. Although this term is still the subject of scientific debate, it has been argued that good surrogate endpoints should be able to predict the causal treatment effect on the true endpoint and many attempts have been made in the literature to evaluate their potential predictive value (Alonso *et al.*, 2017; Buyse *et al.*, 2000; Alonso and Molenberghs, 2007; Gilbert and Hudgens, 2008; Alonso, Van der Elst and Meyvisch, 2016b). When both endpoints are normally distributed, Alonso *et al.* (2015) proposed a methodology to assess the possibility of finding a good surrogate, where "good" was defined in terms of the mean squared error of the prediction of the individual causal treatment effect on the true endpoint, using the individual causal treatment effect on the surrogate. However, when both endpoints are binary, the individual causal treatment effects are categorical variables and the use of the prediction mean squared error becomes less appropriate. In the following, an information-theoretic approach is proposed to assess the plausibility of finding a good surrogate when both endpoints are binary. The method defines "good" in terms of the probability of a prediction error and it is based on Fano's inequality.

In the surrogate marker context, when both endpoints are binary, one wants to predict the individual causal treatment effect on the true endpoint ΔT, based on the individual causal treatment effect on the surrogate ΔS. Let g be a general prediction function and $\widetilde{\Delta T} = g(\Delta S)$ denote the predicted individual causal treatment effect on the true endpoint. The probability of a prediction error can be defined as $P_e = P\left(\widetilde{\Delta T} \neq \Delta T\right)$. Using Fano's inequality it can easily be shown that

$$H\left(P_e\right) + P_e \log\left(|\Delta T| - 1\right) \geq H(\Delta T) - \min\left[H(\Delta T), H(\Delta S)\right] R_H^2 \qquad (2)$$

where $|\Delta T|$ denotes the cardinality of the support of ΔT and the function H is given by $H\left(P_e\right) = -P_e \log\left(P_e\right) - (1 - P_e)\log(1 - P_e)$. Some insight can be gained by considering the function $f\left(P_e\right) = H\left(P_e\right) + P_e \log\left(|\Delta T| - 1\right)$. It can easily be shown that the first derivative df/dP_e takes positive values if and only if $P_e \leq \left(|\Delta T| - 1\right)/|\Delta T|$. In the present setting $|\Delta T|$ is equal to 2 or 3 and, consequently, f will always be an increasing function if $P_e \leq 0.5$. If one fixes an upper bound for the prediction error, say δ, then in practice one would only be interested in surrogate endpoints for which $P_e \leq \delta \leq 0.5$. Due to the monotonicity of f in this region from (2) one finds that if $P_e \leq \delta \leq 0.5$ then

$$R_H^2 \geq \frac{H(\Delta T) - f(\delta)}{\min\left[H(\Delta T), H(\Delta S)\right]} \geq 1 - \frac{f(\delta)}{H(\Delta T)} = R_{HL}^2 \qquad (3)$$

Notice that R_{HL}^2 does not depend on the surrogate. Actually, R_{HL}^2 can be seen as an intrinsic property of the true-endpoint-treatment pair and, in practice, it could be used to assess the plausibility of finding a good surrogate endpoint. In fact, suppose that one would be willing to use a surrogate only if the probability of a prediction error $P\left(\widetilde{\Delta T} \neq \Delta T\right) \leq \delta = 0.3$, i.e., the surrogate may lead to wrong predictions in at most 30% of the cases. Suppose further that for $\delta = 0.3$ the $R_{HL}^2 = 0.9$, then one

would need to find a surrogate endpoint that produces a R^2_H of at least 90% in order to keep the risk of a prediction error smaller than 0.3. Arguably, such a surrogate endpoint may be difficult to find. On the other hand, if for $\delta = 0.3$ a $R^2_{HL} = 0.6$ is obtained then surrogate endpoints with a R^2_H of at least 60% may be capable of keeping the risk of a prediction error smaller than the pre-specified δ. Therefore, although a $R^2_{HL} = 0.6$ does not guarantee that a good surrogate endpoint exists, it certainly makes its existence more plausible.

Some important insights can be obtained from the analysis of the R^2_{HL}. For instance, $\lim_{\delta \to 0} R^2_{HL} = 1$ and, hence, the smaller the risk one is willing to take regarding the prediction error, the more difficult it will be to find a suitable surrogate. Importantly, Fano's inequality and the R^2_{HL} allow the data analyst to visualize the balance between the risk he is willing to take when using the surrogate and the plausibility of finding a surrogate for that level of risk.

3.4 Identifiability issues

When assessing R^2_{HL}, one faces the problem that neither the distribution of ΔT nor its entropy are identifiable from the data, without making untestable assumptions about the association between the potential outcomes for the true endpoint θ_T. For instance, if monotonicity is assumed for the true endpoint $P(T_0 > T_1) = \pi^T_{10} = 0$ (or equivalently $\theta_T = \infty$), then the distribution of ΔT becomes identifiable and, hence, $H(\Delta T)$ and R^2_{HL} become identifiable as well. However, the use of identifiability conditions raises some practical problems. In fact, often there is not enough subject specific knowledge to assess the validity of the identifiability assumptions and, in general, they can be neither proven nor disproven based on the data.

Therefore, a sensitivity analysis may be a more meaningful strategy to handle

unidentifiability in this type of situations. Basically, one would like to study the behavior of R^2_{HL} across all scenarios compatible with the data at hand. To that end, notice that the entropy of the individual causal treatment effect on the true endpoint $H(\Delta T)$ can be written as a function of π^T_{10}. Indeed,

$$H(\pi^T_{10}) = \pi^T_{10}\log\pi^T_{10} + (1-\lambda_m-2\pi^T_{10})\log(1-\lambda_m-2\pi^T_{10}) + (\pi^T_{10}+\lambda_m)\log(\pi^T_{10}+\lambda_m)$$

with π^T_{10} in $[0, \min(\pi^T_{1\cdot}, \pi^T_{\cdot 0})]$ and $\lambda_m = \pi^T_{\cdot 1} - \pi^T_{1\cdot}$. Under SUTVA, if one further assumes that the treatment assignment mechanism is independent of the potential outcomes, then λ_m becomes the identifiable expected causal treatment effect typically estimated in clinical trials. Substituting $\pi^T_{1\cdot}$ and $\pi^T_{\cdot 0}$ by their maximum likelihood estimates and plugging $H(\pi^T_{10})$ into the right side of the last inequality in (3), one obtains R^2_{HL} as a function of π^T_{10}. A plot of this function on the interval $[0, \min(\pi^T_{1\cdot}, \pi^T_{\cdot 0})]$ would allow to assess the behavior of the R^2_{HL} across all scenarios compatible with the data.

The previous approach can also naturally incorporate expert opinion in cases where it is available. In fact, π^T_{10} has a clear clinical interpretation, i.e., it quantifies the probability that a patient would have a better response under the control than under the new treatment. In some scenarios subject-specific knowledge could be used to define clinically meaningful bounds for this probability, for instance, doctors may assess that $0.1 \leq \pi^T_{10} \leq 0.2$. Analyzing the behavior of R^2_{HL} on this interval, one could evaluate the plausibility of finding a good surrogate endpoint, using only the values of R^2_{HL} that are in agreement with the clinically meaningful values of π^T_{10}.

Although appealing, the previous implementation of the sensitivity analysis does not take into account the sampling variability in the estimates of $\pi^T_{1\cdot}$ and $\pi^T_{\cdot 0}$. To solve this problem, along the lines presented in Alonso *et al.* (2016a), the following

sampling algorithm can then be used

From $k = 1$ until M do

1. Uniformly sample a value for $\pi^T_{1\cdot}$ and $\pi^T_{\cdot 0}$ from their corresponding 95% confidence intervals $I_{95}(\pi^T_{1\cdot})$, $I_{95}(\pi^T_{\cdot 0})$
2. Based on the $\pi^T_{1\cdot}$ and $\pi^T_{\cdot 0}$ sampled in step 1, uniformly sample a value for π^T_{10} on the interval $[0, \min(\pi^T_{1\cdot}, \pi^T_{\cdot 0})]$
3. Based on the π^T_{10} sampled in step 2 and the $\pi^T_{1\cdot}$ and $\pi^T_{\cdot 0}$ sampled in step 1 compute all the others $\pi^T_{ij} = P(T_0 = i, T_1 = j)$
4. Based on the distribution of $\boldsymbol{Y}_T$ obtained in step 3 calculate the distribution of ΔT using the expressions $\pi^{\Delta T}_{-1} = \pi^T_{10}$, $\pi^{\Delta T}_0 = \pi^T_{00} + \pi^T_{11}$ and $\pi^{\Delta T}_1 = \pi^T_{01}$
5. Based on step 4 calculate the $H(\Delta T)$ as

$$H(\Delta T) = \pi^T_{10} \log \pi^T_{10} + (\pi^T_{00} + \pi^T_{11}) \log(\pi^T_{00} + \pi^T_{11}) + (\pi^T_{10}) \log(\pi^T_{10})$$

6. Compute R^2_{HL}

This algorithm produces a frequency distribution for R^2_{HL}. This distribution takes into account both the uncertainty induced by the lack of identifiability of the distribution of $\boldsymbol{Y}_T$ and the sampling variability in the estimates of $\pi^T_{1\cdot}$ and $\pi^T_{\cdot 0}$. Basically, this frequency distribution characterizes the behavior of R^2_{HL} across all settings compatible with the data, i.e., all settings compatible with $\boldsymbol{\pi}_T$. Given that all points in $[0, \min(\pi^T_{1\cdot}, \pi^T_{\cdot 0})]$ are equally compatible with the data at hand, the use of a uniform distribution to sample π^T_{10} is the most natural choice. The behavior of R^2_{HL} can then be visualized using graphical techniques like histograms and/or summarized using measures of central tendency like, for instance, the mean and the median, among others.

As stated before, using a uniform sampling scheme on the confidence intervals associated with $\pi_{1\cdot}^T$ and $\pi_{\cdot 0}^T$ is intuitively appealing, however, there is a level of arbitrariness in this choice. Alternatively, one could use a standard normal distribution centered at $\hat{\pi}_{1\cdot}^T$ and $\hat{\pi}_{\cdot 0}^T$ and with standard deviations equal to the associated standard errors. A beta distribution obtained from a flat or Jeffreys prior on these probabilities will also make theoretical and practical sense. Studying the performance of these alternative methods via simulations and real examples goes beyond the scope of these manuscript and will be the focus of future research.

In the following section the previous ideas will be applied to analyze the likelihood of finding a valid surrogate endpoint in the context of glaucoma research.

4 Collaborative Initial Glaucoma Treatment Study

A practical limitation often encountered when validating surrogate endpoints is the lack of user friendly software to carry out the analysis. All the analyses presented in this manuscript can be carried out using the R package *Surrogate*, freely available at *CRAN*. For conciseness, in the following only a summary of the main results is provided and no reference to the software is made. However, in the Supplementary Materials accompanying the paper (available from the authors) a more detailed analysis is provided and the implementation in R discussed.

The Collaborative Initial Glaucoma Treatment Study (CIGTS) was a randomized clinical trial comparing the efficacy of surgery versus a conventional therapy in the treatment of patients suffering from glaucoma. The study included 228 patients of whom 102 received the new treatment and 126 the conventional therapy. 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 (Musch *et al.*, 1999; Li et al., 2011).

Alonso *et al.* (2016a) carried out a detailed analysis of these data and concluded that IOP after 1 year seemed to be a poor surrogate for IOP after 8 years. However, the question remains regarding the existence of a good surrogate endpoint for IOP after 8 years in the first place. First of all, one needs to define what is understood by a good surrogate endpoint in terms of prediction error. We considered three upper bounds $\delta = 0.05, 0.1$ and 0.2, for the probability of a prediction error $P\left(\widetilde{\Delta T} \neq \Delta T\right) \leq \delta$, where $\widetilde{\Delta T} = g(\Delta S)$ and g a general prediction function. For each of these values, the algorithm described in section 3.3 was applied and the corresponding frequency distributions of R^2_{HL} were obtained. Figure 1 (left) summarizes the main results when monotonicity is not assumed. Expectedly, $\delta = 0.05$ led to higher values of R^2_{HL} and, consequently, large values of R^2_H are needed to achieve this pre-specified risk level. In fact, when $\delta = 0.05$, $R^2_{HL} \geq 0.63$ for 80% of the generated $\boldsymbol{Y}_T$ distributions. Actually, the mean and median of R^2_{HL} equaled 0.69 and 0.71 in this scenario, respectively. Clearly, if one wants to keep the risk of a prediction error smaller than 5% then one may need to find a surrogate endpoint which produces a large individual causal association. Notice that a large R^2_H implies an almost deterministic relationship between ΔT and ΔS. Therefore, although not impossible, common sense suggests that finding a surrogate that produces a high ICA will likely be a more daunting task than finding a surrogate that produces a moderate or small ICA.

Let us now consider $\delta = 0.1$, i.e., the surrogate may lead to wrong predictions in at most 10% of the cases. This increased risk is translated into smaller values of R^2_{HL}. Indeed, $R^2_{HL} \leq 0.58$ for 80% of the generated $\boldsymbol{Y}_T$ distributions and $R^2_{HL} \leq 0.61$

for all of them. The mean and median of R^2_{HL} equaled 0.47 and 0.51, respectively. Even though the previous analysis does not guarantee the existence of a good binary surrogate endpoint in this scenario, it does suggest that finding a surrogate that achieved the pre-specified risk level may be more likely in this setting than in the previous one. Ultimately, the data analyst will have to find a balance between the risk level he is willing to accept and the likelihood of finding a good surrogate for that level of risk.

It is important to point out that, in some cases, large values of δ produced negative values of R^2_{HL} and Fano's inequality led to the trivial inequality $R^2_H \geq 0$. This becomes apparent in the peaks observed at zero in the frequency distributions obtained for $\delta = 0.2$. In these cases the results of the analysis are clearly uninformative. However, at an intuitive level, one may expect that if the prediction error comes close to 50% (the tossing coin scenario) then almost any surrogate will be capable of reaching this risk level.

Figure 1 (right) displays the relationship between R^2_{HL} and π^T_{10} for the different δ values. It is important to point out that, unlike the previous analysis, this graph does not take into account the sampling variability in the estimates of $\pi^T_{1\cdot}$ and $\pi^T_{\cdot 0}$. Nonetheless, the results are rather similar, i.e., the smaller the risk (δ) the larger the ICA. For $\delta = 0.05$ the ICA ranges from about 0.5 for smaller values of π^T_{10} till values close to 0.8 when π^T_{10} approaches 0.20. On the other hand for $\delta = 0.1$ the ICA is always smaller than 0.4 irrespectively of the value of π^T_{10}.

The previous analysis suggests that finding a binary surrogate endpoint with a probability of prediction error under 5% may be a challenging task, however, if a 10% risk is considered acceptable then the chances of success may be substantially larger. Obviously, other clinical considerations need to be brought into the discussion

as well, but the previous analysis of R^2_{HL} clearly offers a useful quantitative element for the decision making process.

5 Conclusions

Using information-theoretic elements a procedure was introduced to assess the likelihood of finding a good surrogate when both endpoints are binary. The method relates the probability of a prediction error with the ICA, i.e., the association between the individual causal treatment effects. Measures of association, like the ICA, are easy to interpret and data analysts have a lot of experience working with them. Basically, the methodology gives a quantitative framework to evaluate the relationship between the risk one is willing to take when using a surrogate endpoint and the likelihood of finding a surrogate that achieved that level of risk. Obviously, it is a process that will require not only statistical considerations but also clinical and practical elements will often be taken into account. In general, the assessment of the possibility of finding a good surrogate will not be exempt of a certain degree of subjectivity and, hence, the availability of quantitative elements, like the R^2_{HL}, may be of great valuable during the decision making process.

The need for such a methodology is further motivated by the fact that the evaluation of surrogate endpoints is a challenging task. For instance, the validity of a surrogate for a given true endpoint is treatment dependent. In fact, strictly speaking, an outcome like progression free survival (PFS) would need to be evaluated as a putative surrogate for overall survival (OS) for every new treatment. In addition, the evaluation exercise frequently requires large amount of data. Actually, the most widely accepted methodology for the evaluation of surrogate endpoints, the so-called, meta-analytic approach requires the use of information on both the true and

surrogate endpoint from several clinical trials. Methods that need less information are available, but they are often predicated on untestable assumptions.

In general, even when validated surrogates are used, predicting the treatment effect on the true endpoint, based on the treatment effect on a surrogate, will always convey certain level of risk. That is the reason why regulatory agencies from around the globe have framed the use of surrogate endpoints in the drug discovery process by implementing various provisions and policies. For example, in the United States, there are mechanisms available for accelerated approval based on surrogate endpoints, in order to reduce the time to review an application for indications with no known effective therapy and for providing access to patients for non-approved drugs. Accelerated approval (sometimes referred to as "conditional approval" or "Subpart H") refers to an acceleration of the overall development plan by allowing submission of an application, and if approved, marketing of a drug on the basis of surrogate endpoints while further studies demonstrating direct patient benefit are underway. Accelerated approval is limited to diseases where no effective therapies exist and is based on a surrogate endpoint likely to predict clinical benefit. The implementation of these "conditional approval" policies tries to reduce the potentially harmful consequences of using a surrogate endpoints in the confirmatory stage, while taking advantage of their capacity to accelerate the overall development plan.

Assessing the likelihood of finding a good surrogate in other important settings like, for instance, the continuous-binary, survival-survival, among others, need to be studied as well. They will be the subject of future research.

References

Alonso, A. and Molenberghs, G. (2007). Surrogate marker evaluation from an information theoretic perspective. *Biometrics*, **63**, 180–186.

Alonso, A., Van der Elst, W., Molenberghs, G., Buyse, M. and Burzykowski, T. (2015). On the relationship between the causal-inference and meta-analytic paradigms for the validation of surrogate endpoints. *Biometrics*, **70**, 15–24.

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.

Alonso A and Van der Elst, W., and Meyvisch, P. (2016). Assessing a surrogate predictive value: A causal inference approach. *Statistics in Medicine* **36, 7**, 1083–1098.

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, ISBN 9781482249361 - CAT# K23717.

Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D. and Geys, H. (2000). The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics* **1**, 49–67.

Buyse, M., Sargent, D. J., Grothey, A., Matheson, A. and de Gramont, A. (2010). Biomarkers and surrogate end points–the challenge of statistical validation. *Nat. Rev. Clin. Oncol.* **7**, 309–317.

Elliott, M.R., Li, Y., Taylor, J.M.G. (2013). Accommodating missingness when assessing surrogacy via principal stratification. *Clinical Trials* **10**, 363–377.

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

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

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

Li, Y., Taylor. J.M.G., and Elliott M.R. (2010). A Bayesian approach to surrogacy assessment using principal stratification in clinical trials. *Biometrics* **58**, 21-29.

Li, Y., Taylor, J.M.G., Elliott, M.R., and Sargent, D.R. (2011). Causal assessment of surrogacy in a meta-analysis of colorectal clinical trials. em Biostatistics **12**, 478–492.

Musch, D.C., Lichter, P.,R., Guire, K.,E., Standardi, C.,L., and CIGTS Investigators. (1999). The Collaborative Initial Glaucoma Treatment Study (CIGTS): Study design, methods, and baseline characteristics of enrolled patients. *Ophthalmology* **106**, 653–662.

Micheel C.M., Ball J.R (2010) Evaluation of Biomarkers and Surrogate Endpoints in Chronic Disease. Washington, DC: National Academies Press. http://www.iom.edu/Reports/2010/Evaluation-of-Biomarkers-and-Surrogate-Endpoints-in-Chronic-Disease.aspx. Accessed October 31, 2013.

Plackett R.L.(1965) A class of bivariate distributions. *Journal of the American Statistical Association*, **60**, 516–522.

Rubin, D. B. (1980). Randomization analysis of experimental-data the Fisher randomziation testâĂŤcomment. *Journal of the American Statistical Association* **75**, 591–593.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal* **27**, 379–423 and 623–656.

Writing Group for the Woman's Health Initiative Investigators (2002). Risks and benefits of estrogen plus progestin in healthy postmenopausal women. *Journal of the American Medical Association*, **288**, 321–333.

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

		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

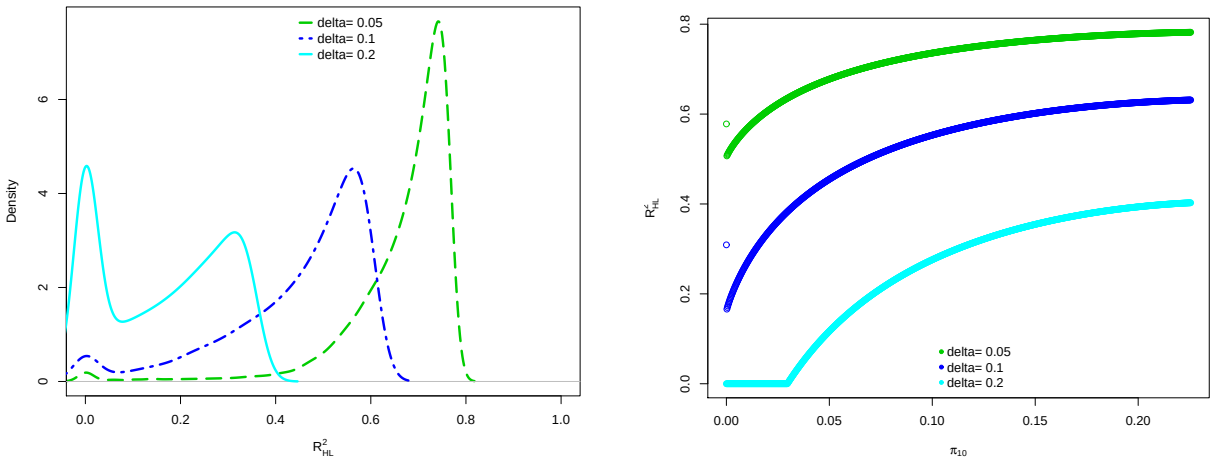

Figure 1: R^2_{HL} when monotonicity is not assumed (left) and R^2_{HL} versus π^T_{10} (right).