A reflection on the causal interpretation of individual-level surrogacy

Ariel Alonso^{*,1} **Wim Van der Elst²**

Geert Molenberghs³ **Alvaro J. Florez¹**

¹, I-BioStat, Katholieke Universiteit 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

At the beginning of the 21st century, a new paradigm was introduced for the evaluation of surrogate endpoints based on meta-analysis. In this paradigm, the putative surrogate is assessed at two different levels, the so-called, trial and individual level. Trial level surrogacy is defined as the association between the expected causal treatment effects across different trials populations, whereas the individual level is defined as the association between the surrogate and true endpoints, after adjusting by trial and treatment. It has been argued that the individual level surrogacy does not have a causal interpretation and, consequently, it is a poor metric of surrogacy. In the present work, an alternative definition of individual level surrogacy is introduced based on individual causal treatment effects. In addition, using the maximum entropy principle, a direct link between the individual level surrogacy, as defined in the meta-analytic approach, and the newly proposed definition is established. This new perspective sets the individual level surrogacy in a more coherent framework with respect to the trial level and bridges the two main schools of thought in this domain, namely, the causal inference and meta-analytic schools.

Keywords: Causal inference, Meta-analytic approach, Maximum entropy, Surrogate endpoints.

1 Introduction

At the beginning of the 21st century, a new paradigm was proposed for the evaluation of surrogate endpoints in a meta-analytic framework. In this paradigm, the putative surrogate is assessed at two different levels, the so-called, trial and individual level. At the trial level, expected causal treatment effects, i.e., the averages of the individual causal effects across all patients within the trial populations, are used to carry out the evaluation exercise. Basically, trial level surrogacy is defined as the association between the expected causal treatment effects. Unlike the trial level, individual level surrogacy is not based on causal treatment effects, but rather defined as the association between the surrogate and true endpoints, after adjusting for trial and treatment (Gail et al., 2000; Buyse et al., 2000). It has been argued that this definition of individual level surrogacy does not have a direct causal interpretation and, therefore, it suffers from the same serious drawbacks as Prentice's main criterion (Frangakis and Rubin, 2002). In addition, it has now been clearly established that, although appealing, measures of association between the surrogate and true endpoints have limited value as metrics of surrogacy (Fleming and DeMets, 1996; Burzykowski *et al.*, 2005). Taking into account these limitations, in the present work an alternative definition of individual level surrogacy is introduced based on individual causal treatment effects. Essentially, the individual level surrogacy is quantified using the individual causal association (ICA), i.e., the association between the individual causal effects. This new definition has a direct causal interpretation and a natural, logical connection with the definition used at the trial level. Furthermore, using the maximum entropy principle (MEP) (Jaynes, 1985), it is shown that the individual level surrogacy defined in the classical meta-analytic approach is the maximum entropy (ME) value of the ICA.

According to the maximum entropy principle, if nothing is known about a distribu-

tion except that it belongs to a certain class, then the distribution with the largest entropy should be chosen as the least-informative default. The MEP can be traced back to the end of the nineteenth century, when the principles of statistical mechanics and the second law of thermodynamics were developed. In the second half of the twenty century, Edwin T. Jaynes extended the principle to more general scientific problems (Jaynes, 1985, 2003) and since then it has been successfully applied in many scientific domains, like radio astronomy (Bridle *et al.*, 1998), modeling turbulence in fluids and plasmas (Montgomery, 1996) and modeling proteins sequences (MacKay, 1996), among others.

The newly proposed definition sets the individual level surrogacy in a more coherent context with respect to the trial level and bridges the two main paradigms for the evaluation of surrogate endpoints, the so-called causal inference and meta-analytic paradigms. Firstly, the relationship between the adjusted association and the ME value of the ICA is established in the single trial setting. Secondly, the results from the single trial setting are used to redefine the individual level surrogacy in the meta-analytic framework and to explore its connection with the classical definition.

In Section 2, the causal-inference model underlying the validation strategy is introduced. In Section 3, the MEP is presented and used to explore the relationship between the adjusted association and ICA, in the single trial setting. The meta-analytic approach to the validation of surrogate endpoints is reviewed in Section 4, where a new definition of individual level surrogacy is also given. Furthermore, the underlying link between the classical and new definition is studied using the MEP. The identifiability problems associated with the ICA are studied in Section 5 where approximate identifiable bounds are proposed and studied via simulations. The case study is introduced and analyzed in Section 6 and some final comments are given in Section 7.

2 Causal-inference model

As previously stated, we will start by considering the single-trial setting (STS). In this setting, the trial's protocol precisely defines the population in which the validation exercise is framed and it will further be assumed that only two treatments are under evaluation ($Z = 0/1$).

Rubin's model for causal inference states that each patient has a four-dimensional vector of potential outcomes $\mathbf{Y} = (T_0, T_1, S_0, S_1)'$. The components T_1 , S_1 , T_0 and S_0 are potential outcomes in that they represent the outcomes for the true and surrogate endpoints of an individual had he received the treatment or control, respectively. The so-called fundamental problem of causal inference asserts that, often in practice, only two of these four potential outcomes are observed (Holland, 1986).

The vector of individual causal treatment effects on the surrogate and true endpoint can be defined as $\Delta = (\Delta T, \Delta S)'$, where $\Delta T = T_1 - T_0$ and $\Delta S = S_1 - S_0$. Note that, due to the fundamental problem of causal inference, neither ΔT nor ΔS are identifiable from the data.

The expected causal effects of the treatment on the surrogate and true endpoint, in the population under study, are defined as $\alpha = E(\Delta S)$ and $\beta = E(\Delta T)$. Unlike the individual causal effects, the expected causal effects are identifiable from the data under fairly general conditions. In fact, Rosenbaum and Rubin (1983) provided three identifiability conditions under which it is possible to obtain consistent estimators of the expected causal treatment effects. If Y denotes the response of interest and Y_z the potential outcome associated with $Z = z$ then the three identifiability conditions are: 1) Consistency: If $Z = z$ for a given subject then $Y_z = Y$ for that subject, 2) Conditional exchangeability: This condition essentially states that

there are not unmeasured confounding given data on baseline covariates L , that is, $Y_z \perp Z|L = l$ for each possible value z of Z and l of L and 3) Positivity: If $f_L(l) \neq 0$ then $f_{Z|L}(z|l) > 0$. It can easily be shown that in randomized clinical trials all condition hold and the expected causal treatment effects can be identified as $\beta = E(T|Z = 1) - E(T|Z = 0)$ and $\alpha = E(S|Z = 1) - E(S|Z = 0)$, with the expected values estimated using the sample means of T and S in the control and treated groups, respectively. The ICA, proposed by Alonso *et al.* (2015), and the methods based on it that will be proposed in the following sections, are based only on individual causal treatment effects and are valid if consistency holds, i.e., they could also be applied to observational data.

Although less relevant for the estimation of the expected causal effects, the distribution of the vector of potential outcomes plays an important role in the surrogate validation context. In the following sections, the focus will be on a setting that has already received a lot of attention in the literature, namely, the scenario in which both the surrogate and true endpoint are continuous and normally distributed outcomes (Buyse and Molenberghs, 1998; Buyse et al., 2000; Gail et al., 2000; Conlon, Taylor, and Elliott, 2013; Alonso *et al.*, 2017). Therefore, along the lines of Conlon, Taylor, and Elliott (2013) and Alonso *et al.* (2015), it will be further assumed that $\mathbf{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}' = (\mu_{T0}, \mu_{T1}, \mu_{S0}, \mu_{S1})$ and

$$\boldsymbol{\Sigma} = \left(\begin{array}{cc|cc} \sigma_{T0} & \sigma_{T0T1} & \sigma_{T0S0} & \sigma_{T0S1} \\ \sigma_{T0T1} & \sigma_{T1} & \sigma_{T1S0} & \sigma_{T1S1} \\ \hline \sigma_{T0S0} & \sigma_{T1S0} & \sigma_{S0} & \sigma_{S0S1} \\ \sigma_{T0S1} & \sigma_{T1S1} & \sigma_{S0S1} & \sigma_{S1} \end{array} \right).$$

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). Following this idea, Alonso *et al.* (2015) proposed to assess surrogacy using the so-called individual causal association. When both endpoints are continuous and normally distributed the ICA is quantified by the Pearson correlation coefficient $\rho_{\Delta} = \text{corr}(\Delta T, \Delta S)$. Given the aforementioned distributional assumptions one has

$$\Delta = \mathbf{A}\mathbf{Y} = \begin{pmatrix} T_1 - T_0 \\ S_1 - S_0 \end{pmatrix} \sim N(\boldsymbol{\mu}_{\Delta}, \boldsymbol{\Sigma}_{\Delta}), \quad (1)$$

where $\boldsymbol{\Sigma}_{\Delta} = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$, $\boldsymbol{\mu}_{\Delta} = (\beta, \alpha)'$ and $\mathbf{A}$ the corresponding contrast matrix. After some algebraic calculations it can easily be shown that

$$\rho_{\Delta} = \frac{\sqrt{\sigma_{T0}\sigma_{S0}}\rho_{T0S0} + \sqrt{\sigma_{T1}\sigma_{S1}}\rho_{T1S1} - \sqrt{\sigma_{T1}\sigma_{S0}}\rho_{T1S0} - \sqrt{\sigma_{T0}\sigma_{S1}}\rho_{T0S1}}{\sqrt{(\sigma_{T0} + \sigma_{T1} - 2\sqrt{\sigma_{T0}\sigma_{T1}}\rho_{T0T1})(\sigma_{S0} + \sigma_{S1} - 2\sqrt{\sigma_{S0}\sigma_{S1}}\rho_{S0S1})}}, \quad (2)$$

where ρ_{XY} denotes the correlation between the potential outcomes X and Y . Notice that, under consistency, all the variances and the correlations ρ_{T0S0} , ρ_{T1S1} , are identifiable and can be estimated using data from the treated and control group. However, the other correlations are not identifiable and, hence, ρ_{Δ} cannot be estimated. The ICA can 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. In fact, it takes value ± 1 if there is a deterministic relationship between ΔT and ΔS and, therefore, knowing what the treatment does to the surrogate on a specific patient brings full information about what it will do to his true endpoint. In addition, when the ICA equals zero both treatment effects are independent and knowing what the treatment does to the surrogate on a specific patient brings no information

whatsoever about what it will do to his true endpoint.

If one further assumes that $\sigma_{T0} = \sigma_{T1} = \sigma_T$ and $\sigma_{S0} = \sigma_{S1} = \sigma_S$, i.e., the variability of the true and surrogate endpoint is constant across the two treatment conditions, then

$$\rho_{\Delta} = \frac{\rho_{T0S0} + \rho_{T1S1} - \rho_{T1S0} - \rho_{T0S1}}{2\sqrt{(1 - \rho_{T0T1})(1 - \rho_{S0S1})}}. \quad (3)$$

The previous assumption of homoscedasticity is inherent to many statistical techniques like linear regression and analysis of variance, and it is testable using the observable data. In the rest of the manuscript this assumption will be adopted to simplify the algebraic calculations. However, the derived conclusions will be also valid when all variances are different.

3 Maximum entropy principle

The MEP states that, when making inferences based on incomplete information, one should use the probability distribution that has the maximum entropy permitted by the data at hand. Shannon (1948) introduced the concept of entropy to quantify the “epistemic” uncertainty or lack of knowledge implied by a distribution. Information and entropy are opposite concepts, i.e., to gain information is to loose uncertainty by the same amount and, hence, their formal definitions differ only in sign. If Y is a discrete random variable taking values $\{y_1, y_2, \dots, y_m\}$ with probability function $P(Y = y_i) = p_i$, then the entropy of Y is defined as $H(Y) = -E_Y[\log P(Y)] = -\sum_i p_i \log p_i$. For continuous outcomes the concept of entropy needs to be extended and the so-called differential entropy is used instead. For a continuous random variable Y with density $f_Y(y)$ and support Ψ the differential entropy $h_d(Y)$ is defined

as

$$h_d(Y) = -E_Y[\log f_Y(Y)] = - \int_{\Psi} f_Y(y) \log f_Y(y) dy.$$

Differential entropy enjoys some but not all properties of entropy, e.g., it can be infinitely large, negative, or positive, and it is coordinate dependent. If $\mathbf{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then

$$h_d(\mathbf{Y}) = \frac{1}{2} \ln \left((2\pi e)^k |\boldsymbol{\Sigma}| \right), \quad (4)$$

where k denotes the dimension of $\mathbf{Y}$ and, therefore, in the present context, $k = 4$. From (4), it follows that maximizing the entropy is equivalent with maximizing the determinant of the matrix $\boldsymbol{\Sigma}$. However, not all the entries in $\boldsymbol{\Sigma}$ are identifiable. Indeed, due to the fundamental problem of causal inference, the parameters $\mathcal{L} = \{\sigma_{T_0T_1}, \sigma_{T_0S_1}, \sigma_{T_1S_0}, \sigma_{S_0S_1}\}$ cannot be estimated from the data. This lack of identifiability gives rise to a matrix completion problem. The covariance matrix $\boldsymbol{\Sigma}$ is a partial matrix, i.e., a rectangular array in which certain entries are specified, while the remaining entries are unspecified and free to be chosen from an indicated set, such as the real numbers. In the present example, the set $\mathcal{L}$ contains all the unspecified elements of $\boldsymbol{\Sigma}$. A choice of values for the unspecified entries results in a completion of the partial matrix, and a matrix completion problem asks whether there is a completion with some desired properties like, for instance, a positive definite matrix completion that maximizes (4) (Glunt et al., 1999; Grone et al., 1984). The following lemma characterizes the solution of maximizing the entropy (4) over $\mathcal{L}$.

Lemma 1 *Among all positive definite completions on $\mathcal{L}$, there is a unique one with maximum determinant. Furthermore, the determinant maximizing completion is characterized by having all the elements in $\mathcal{L}$ equal to zero.*

The proof of Lemma 1 is provided in the Supplementary Materials accompanying this paper. The motivation behind the use of the MEP is twofold: firstly, maximizing entropy minimizes the amount of prior information built into the distribution; and secondly, empirical evidence shows that most physical systems tend to move towards maximal entropy configurations over time.

In the STS, Buyse and Molenberghs (1998) proposed to evaluate a surrogate endpoint using the so-called relative effect and adjusted association (AA). The relative effect is defined as $RE = \beta/\alpha$, whereas the adjusted association is given by the correlation between the surrogate and true endpoints after adjusting by treatment $\gamma = \text{corr}(S, T|Z)$. In the STS the validity of the surrogate is studied in a fixed, well defined population and, consequently, only one pair of constant expected causal treatment effects (α, β) is available. This lack of replication/variability limits the value of RE as a metric of surrogacy. Unlike for the expected causal treatment effects, in the STS one does have replication/variability at the level of the individual patient and, therefore, at the level of the individual causal treatment effects. The ICA quantifies the association at this level, i.e., the association between ΔT and ΔS . One may wonder if there is any relationship between the AA and the ICA. The following lemma answers this question.

Lemma 2 *If the causal inference model introduced in Section 2 is valid, then the maximum entropy value of the individual causal association, i.e., the ICA obtained from the maximum entropy distribution of the potential outcomes, is given by*

$$\rho_{\Delta} = \frac{\rho_{T0S0} + \rho_{T1S1}}{2}.$$

If one further assumes that $\rho_{T0S0} = \rho_{T1S1} = \gamma$ then $\rho_{\Delta} = \gamma$.

Lemma 2 shows that if no additional information on the parameters in $\mathcal{L}$ is available, then the AA offers the most general assessment of the ICA, i.e., the assessment based on the distribution with minimal informational content.

4 Multi-trial setting

The meta-analytic approach offered an alternative method to carry out the validation of surrogate endpoints. In this approach, it is assumed that the new treatment is evaluated in a sequence of clinical trials and, although similar, these trials target slightly different populations (due to differences in the inclusion/exclusion criteria, protocol and/or differences between countries where the trials take place, among other reasons) and estimate potentially different expected causal treatment effects for the surrogate and true endpoint. The main question addressed in this paradigm is: How much information does the expected causal treatment effect on the surrogate convey on the expected causal treatment effect on the true endpoint? To that end, let us assume that data from $i = 1, \dots, N$ clinical trials are available, and let n_i denote the number of patients included in the i th trial. Furthermore, let T_{ij} and S_{ij} denote the true and surrogate endpoints for patient j in trial i , respectively, and Z_{ij} be the indicator variable for the new treatment. The random treatment allocation in a clinical trial context naturally leads to the following bivariate model

$$\begin{cases} T_{ij} = \mu_{Ti} + \beta_i Z_{ij} + \varepsilon_{Tij}, \\ S_{ij} = \mu_{Si} + \alpha_i Z_{ij} + \varepsilon_{Sij}, \end{cases} \quad (5)$$

where μ_{Ti} and μ_{Si} are trial-specific intercepts, β_i and α_i are trial-specific expected causal treatment effects and ε_{Tij} and ε_{Sij} are correlated error terms, assumed to

follow a bivariate normal distribution with mean zero and variance-covariance matrix $\Sigma_I = \begin{pmatrix} \sigma_T & \sigma_{TS} \\ \sigma_{TS} & \sigma_S \end{pmatrix}$. Although in the present work a frequentist approach is taken, a fully Bayesian hierarchical version of model (5) has also been proposed in the literature (Shkedy and Barbosa, 2005).

Adjusting for prognostic covariates in (5) can lead to smaller standard errors for the expected causal treatment effects (Kahan et al., 2014). However, such an adjustment will substantially affect the interpretation of these effects. In fact, after adjustment, the expected causal treatment effects (α_i, β_i) will be conditional averages, i.e., the averages of the individual causal treatment effects across the sub-populations defined by the covariates and not across the entire population. Consequently, the results of the validation exercise will only be useful for new trials in which the same adjustment is carried out (same prognostic covariates are measured). One should also keep in mind that regulatory authorities are often interested in the expected causal treatment effects across the entire population for regulatory purposes. Therefore, although some potential benefits may be obtained from adjusting, there are also some caveats that would need to be considered before such a decision is made.

At the trial level, it is further assumed that the vector of trial-specific parameters $(\mu_{Si}, \mu_{Ti}, \alpha_i, \beta_i)'$ follows a normal distribution and, consequently, the vector of expected causal effects $\boldsymbol{\mu}_{\Delta i} = (\alpha_i, \beta_i)'$ is also normally distributed with mean $\bar{\boldsymbol{\mu}}_{\Delta}$ and variance-covariance matrix $\mathbf{D}_{\Delta} = \begin{pmatrix} d_{aa} & d_{ab} \\ d_{ab} & d_{bb} \end{pmatrix}$.

The matrix $\mathbf{D}_{\Delta}$ contains all the information about the association between both expected causal treatment effects and based on it, the expected causal association (ECA) can be defined as the correlation coefficient $R_{\text{trial}} = \text{corr}(\alpha_i, \beta_i) = d_{ab}/(\sqrt{d_{aa}d_{bb}})$. The expected causal association can be seen as a special case of the general approach followed by Buyse et al. (2000) to quantify trial level surrogacy. It

is important to point out that these authors used the squared correlation R_{trial}^2 in their evaluation strategy, a slightly different but equivalent quantity. Further, Buyse et al. (2000) completed the validation procedure by quantifying the individual level surrogacy using the correlation coefficient $R_{\text{ind}} = \sigma_{TS} / \sqrt{\sigma_T \sigma_S}$. Note that R_{ind} is the correlation between both endpoints with adjustment for treatment and trial and, therefore, it is a natural extension of the adjusted association.

Given that the trial level surrogacy is basically defined as the association between the expected causal treatment effects, one could alternatively define the individual level surrogacy as the association between the individual causal treatment effects. The trial level surrogacy assesses the validity of the surrogate at the population level, across different populations. In other words, if a surrogate is valid at the trial level, then the expected causal treatment effect on the surrogate in a new population brings information about the expected causal treatment effect on the true endpoint in this population as well. In addition, the ICA assesses the validity of the surrogate at the level of the individual patients, across different patients. In other words, if a surrogate is valid at this individual level (large ICA) then the individual causal treatment effect on the surrogate for a specific patient brings information about the individual causal treatment effect on the true endpoint for the same patient. However, in the classical meta-analytic approach the individual level surrogacy was defined not in terms of the individual causal treatment effects, but rather in terms of the association between the surrogate and true endpoint outcomes. Undoubtedly, the association between S and T has value in itself and one may wonder if R_{ind} is somehow related to this alternative way of defining individual level surrogacy, i.e., if R_{ind} is related to the ICA.

To explore this possible link, let us extend the causal inference model introduced in Section 2 by assuming that $\mathbf{Y}_{ij} \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, where $\mathbf{Y}_{ij}$ is the vector of potential

outcomes for patient j in trial i and $\boldsymbol{\mu}_i$ being normally distributed. This further implies that $\boldsymbol{\Delta}_{ij} \sim N(\boldsymbol{\mu}_{\Delta i}, \boldsymbol{\Sigma}_{\Delta})$ and $\boldsymbol{\mu}_{\Delta i} \sim N(\bar{\boldsymbol{\mu}}_{\Delta}, \boldsymbol{D}_{\Delta})$, where $\boldsymbol{\Delta}_{ij}$ is the vector of individual causal treatment effects for patient j in trial i , $\boldsymbol{\mu}_{\Delta i}$ is the vector of expected causal treatment effects in trial i , and $\boldsymbol{\Sigma}_{\Delta}$ and $\boldsymbol{D}_{\Delta}$ are defined as before. Under these assumptions:

$$\begin{cases} \Delta T_{ij} = \beta_i + \varepsilon_{\Delta T_{ij}}, \\ \Delta S_{ij} = \alpha_i + \varepsilon_{\Delta S_{ij}}, \end{cases} \quad (6)$$

with $(\varepsilon_{\Delta T_{ij}}, \varepsilon_{\Delta S_{ij}})' \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\Delta})$ and $(\varepsilon_{\Delta T_{ij}}, \varepsilon_{\Delta S_{ij}})' \perp (\alpha_i, \beta_i)'$. Based on model (6), one can quantify the individual level surrogacy using the ICA after conditioning by trial and treatment, i.e.,

$$\rho_{\Delta} = \text{corr}(\Delta T_{ij}, \Delta S_{ij} | \alpha_i, \beta_i) = \text{corr}(\varepsilon_{\Delta T_{ij}}, \varepsilon_{\Delta S_{ij}})$$

Note that like $\boldsymbol{\Sigma}_I$ in model (5), $\boldsymbol{\Sigma}_{\Delta}$ is assumed to be constant across trial populations and, consequently, ρ_{Δ} (like R_{ind}) is constant across trials populations as well.

In the previous framework the ICA assesses the validity of the surrogate within each specific trial population, whereas R_{trial} assesses its validity across the different trial populations. Actually, when focusing on one single population, the expected causal treatment effects are constant parameters and the question about how much information α conveys about β becomes meaningless. However, unlike the expected causal effects, the individual causal treatment effects do vary across different individuals in the population and, therefore, the ICA is a more appropriate metric of surrogacy in this context. In addition, if the validation exercise is carried out across different populations then the expected causal treatment effects can be conceptualized as random variables and the question about how much information α_i conveys about β_i becomes meaningful. In such scenario R_{trial} and ρ_{Δ} offer comple-

mentary information regarding the validity of the surrogate across and within trials, respectively. Following the same line of thought used in Lemma 2 leads to

Lemma 3 *Under the aforementioned distributional assumptions, the maximum entropy value of the individual causal association, i.e., the ICA obtained from the maximum entropy distribution of the potential outcomes, is given by:*

$$\rho_{\Delta} = \text{corr}(\Delta T_{ij}, \Delta S_{ij} | \alpha_i, \beta_i) = R_{ind}.$$

Lemma 3 shows a nice connection between the definition of individual level surrogacy offered by the classical meta-analytic approach and the definition proposed in the present work. Essentially, R_{ind} can be seen as the most general assessment of the ICA in this context, i.e., the assessment based on the distribution of the potential outcomes with minimal informational content.

5 Assessing the ICA

Due to the fundamental problem of causal inference, affecting most causal inference models, the ICA is not identifiable (Holland, 1986). A well-founded approach to deal with this lack of identifiability is the implementation of a sensitivity analysis along the lines presented in Alonso *et al.* (2015). In such a sensitivity analysis, the ICA is assessed by uniformly sampling the unidentified correlations in (2) that are compatible with the data at hand. The algorithm outputs a frequency distribution of the values of the ICA compatible with the data. This frequency distribution can be subsequently summarized using histograms and measures of variability and central tendency. Van der Els, Molenberghs, and Alonso (2016) studied the general performance of the algorithm via simulations.

5.1 Approximate identifiable bounds for the ICA

Although the ICA is unidentifiable and, hence, not estimable, approximate identifiable bounds can be constructed for it. In the present session such bounds are introduced and studied via simulations.

Essentially, one is interested in the behavior of the function $ICA: \Gamma_D \rightarrow [0, 1]$ where $\Gamma_D = \left\{ (\sigma_{T0T1}, \sigma_{T0S1}, \sigma_{T1S0}, \sigma_{S0S1}) : \text{so that } \hat{\Sigma} \text{ is positive definite} \right\}$ and $\hat{\Sigma}$ denoting the matrix Σ with the identifiable entries substituted by their estimated values. Given that all variances in Σ are estimable, the previous set could be equivalently rewritten as $\Gamma_D = \left\{ (\rho_{T1S0}, \rho_{T0S1}, \rho_{T0T1}, \rho_{S0S1},) : \text{so that } \hat{\Sigma} \text{ is positive definite} \right\}$. The sensitivity analysis uniformly samples points $(\rho_{T1S0}, \rho_{T0S1}, \rho_{T0T1}, \rho_{S0S1},)$ on Γ_D and computes the ICA for each of them. If a sufficiently large number of points are sampled, then the interval defined by the observed $\min_{\Gamma_D} ICA$ and $\max_{\Gamma_D} ICA$ should contain the true value of the ICA. Therefore, these minimum and maximum observed values could be interpreted as approximate bounds for the individual causal association.

5.2 Simulation study

We assessed the performance of the previous approximate bounds via simulations. It is important to point out that simulation studies have major limitations in this context. In fact, even though simulations are excellent tools to evaluate the performance of estimation and inferential procedures, their utility is limited when the parameters of interest are unidentifiable. Actually, the data do not convey enough information to estimate the parameters consistently and the same will be true for any generated data set. As stated before, untestable assumptions are often made to make inferences about the unidentifiable parameters. In such a situation it is always

possible to design simulation studies where the proposed methodology produces biased results (the data generating mechanism does not satisfy the assumptions) or unbiased results (the data generating mechanism satisfies the assumptions). The aforementioned algorithm is not designed to estimate the unidentifiable ICA, but to offer a sensitivity analysis about its plausible values.

For the simulations, the following correlation matrices for (T_0, T_1, S_0, S_1) were considered:

$$R_1 = \begin{pmatrix} 1 & 0.52 & 0.89 & 0.41 \\ & 1 & 0.80 & -0.16 \\ & & 1 & 0.24 \\ & & & 1 \end{pmatrix}, R_2 = \begin{pmatrix} 1 & 0.79 & 0.89 & 0.16 \\ & 1 & 0.75 & -0.16 \\ & & 1 & 0.03 \\ & & & 1 \end{pmatrix},$$

and

$$R_3 = \begin{pmatrix} 1 & 0.55 & 0.89 & -0.52 \\ & 1 & 0.22 & -0.16 \\ & & 1 & -0.43 \\ & & & 1 \end{pmatrix},$$

with $V(T_0) = 176.78$, $V(T_1) = 11.83$, $V(S_0) = 0.6213$, and $V(S_1) = 0.0021$. R_1 , R_2 , R_3 lead to ICA values of 0.75, 0.85 and 0.95, respectively.

For each correlation matrix, 500 datasets of size 200 were simulated. Using each data set the identifiable bounds for the ICA were calculated by sampling $M = \{100, \dots, 50000\}$ points on Γ_D . Table 1 displays the average min and max ICA obtained from the algorithm, and the percentage of cases in which the true ICA was included in the $[\min, \max]$ interval (coverage).

Table 1: Average [min, max] interval and coverage based on different number of runs of the algorithm and true ICA values.

M	ICA of 0.75		ICA of 0.85		ICA of 0.95	
	[min, max]	coverage	[min, max]	coverage	[min, max]	coverage
100	(0.75, 0.97)	0.50	(0.75, 0.97)	1	(0.75, 0.97)	0.96
200	(0.74, 0.97)	0.56	(0.74, 0.97)	1	(0.74, 0.97)	0.98
500	(0.73, 0.98)	0.63	(0.73, 0.98)	1	(0.73, 0.98)	0.99
1,000	(0.73, 0.98)	0.67	(0.73, 0.98)	1	(0.73, 0.98)	0.99
5,000	(0.72, 0.98)	0.73	(0.72, 0.98)	1	(0.72, 0.98)	1
10,000	(0.72, 0.98)	0.75	(0.72, 0.98)	1	(0.72, 0.98)	1
50,000	(0.72, 0.98)	0.78	(0.72, 0.98)	1	(0.72, 0.98)	1

The coverage for true ICA of 0.85 and 0.95 was large, even with few runs of the algorithm. However, when true ICA equals 0.75, the coverage also increases but much slower and never surpass 80%. It is important to point out that, even though the coverage was low in this setting, the average lower bound of the intervals were always very close to the true ICA value. An examination of the condition numbers for R_1 , R_2 and R_3 , showed that they equaled 88.82, 33.09 and 72.73 respectively. Condition numbers close to 100 are indicative of an almost singular matrix. Note that the worse results were actually obtained for the matrix with the largest condition number, what may indicate some numerical instability. In general, the finding of the simulations seems to suggest that finding appropriate identifiable bounds may require at least 50,000 runs of the algorithm, i.e., at least 50,000 points should be uniformly sampled on Γ_D .

6 Case study

The preceding ideas will be illustrated using a randomized clinical trial in age-related macular degeneration (ARMD). The ARMD data have been analyzed before in the

surrogate validation literature. Although the reiterative use of an example may seem at first a disadvantage, when various frameworks and the results that they yield are juxtaposed, it is useful to re-analyze such data. Like the famous iris data set in classification analysis and many other well known data sets in other domains, the ARMD data have played a very important role in the development of strategies for the evaluation of surrogate endpoints.

Patients suffering from age-related macular degeneration tend to progressively lose vision (Pharmacological Therapy for Macular Degeneration Study Group, 1997). The objective of the trial was to examine the efficacy of interferon- α to treat age-related macular degeneration. The surrogate and true endpoints were defined based on visual acuity. Visual acuity was examined using standardized vision charts that display lines with five letters of decreasing sizes. The patients had to read these letters from top (largest letters) to bottom (smallest letters). Visual acuity was quantified as the total number of letters that were correctly read by a patient.

The main objective of the analysis was to assess whether change in visual acuity 24 weeks after starting the treatment is a good surrogate for the change in visual acuity 52 weeks after the start of the treatment. A total of 84 and 97 patients were enrolled in the placebo and interferon- α treatment conditions, respectively. Patients were distributed across 36 medical centers.

The R package *Surrogate*, freely available at <http://cran.r-project.org/>, was used for carrying out the analysis. For conciseness, in the present section only a summary of the main results is given and no reference to the software is made. However, in the supplementary materials accompanying the manuscript a more comprehensive analysis of the ARMD trial is provided and the use of the R package *Surrogate* is explained in detail.

As expected, the surrogate and true endpoints were moderately correlated in both treatment groups with $\hat{\rho}_{T0S0} = 0.77$ (CI (0.70, 0.82)) and $\hat{\rho}_{T1S1} = 0.71$ (CI (0.63, 0.78)). Although it seems reasonable to assume that the presence of such correlations is a necessary condition for the validity of a surrogate, there is now a wide consensus on the fact that it is not sufficient (Burzykowski *et al.*, 2005; Alonso *et al.*, 2017). Essentially, the use of surrogate endpoints is intended to learn about what the treatment will do to the true endpoint by studying its effect on the surrogate. In this context, these correlations have only limited value.

The implementation of the meta-analytic approach requires the use of information on both the true and surrogate endpoint from several clinical trials. The need for the availability of such information is an important problem hindering the use of the methodology. This lack of multi-trial data is more acute in the early stages of the drug development process, arguably the moment in which surrogate endpoints are needed most. To address this issue it has been suggested to use alternative units of analysis. For instance, in a single trial setting the countries or centers where patients were treated could be used to assess “trial” level surrogacy. The selection of these alternative units has been studied in the literature but it remains a controversial issue and always requires careful thought (Burzykowski *et al.*, 2005). In the ARMD trial the centers where the patients were treated are typically considered as the unit of analysis for the implementation of the meta-analytic approach.

Fitting the full random-effects model given in (5) produced numerical problems. Infinite likelihoods and an ill-conditioned covariance matrices for the random effects are not uncommon when using model (5) and simplifying strategies have been proposed to overcome these issues (Alonso *et al.*, 2017). One such strategy consists of first fitting the bivariate regression model (5) considering the center-specific parameters as fixed effects and, subsequently, estimate the ECA as $\hat{R}_{\text{trial}} = \text{corr}(\hat{\alpha}_i, \hat{\beta}_i)$,

where $(\hat{\alpha}_i, \hat{\beta}_i)$ are the maximum likelihood estimators obtained at the first stage. Typically, the different contributions of the trials/units are taken into account when estimating the ECA by using weights proportional to their sizes, i.e., the number of patients enrolled in them.

The implementation of the previous procedure leads to an estimate of the trial level surrogacy $\hat{R}_{trial} = 0.84$ (95% CI: [0.70; 0.92]), indicating a moderately strong association between the expected causal treatment effects on S and T , respectively. Therefore, the average effect of the treatment on the surrogate in a given population (α_i) does provide information about the average effect of the treatment on the true endpoint in the same population (β_i). The graph in the top left corner of Figure 1 offers a graphical representation of the ECA. Similarly, to assess the validity of the surrogate at the level of the individual patients one needs to evaluate if the individual causal treatment effect on the surrogate for a patient (ΔS) conveys information on the individual causal effect of the treatment on the true endpoint for the same patient (ΔT). As discussed in the previous sections, the ICA metric can be used for that propose.

As previously stated, the ICA is not identifiable and to tackle this problem the algorithm introduced in Section 5 was applied. The algorithm outputs a frequency distribution of the values of the ICA compatible with the data at hand (see top right corner of Figure 1). The mean $\rho_{\Delta} = 0.73$ and mode 0.78 indicate that the strength of the association between ΔS and ΔT is moderate on average. In addition, 80% and 50% of the ICA values were larger than 0.62 and 0.76, respectively. However, the variability of the ρ_{Δ} values was large with $SD = 0.19$ and the identifiable bounds introduced in Section 5.1 equaled $[-0.25; 0.99]$. This finding demonstrates that the unverifiable assumptions made for the unidentifiable correlations have a large impact on the results. The MEP-ICA value, i.e., the value of the ICA based on the

maximum entropy distribution of the potential outcomes equaled 0.74. Therefore, the assessment of the ICA based on the distribution of $\mathbf{Y}$ with the smallest informational content also indicates a moderate association between ΔS and ΔT . For comparison this value is given by a vertical line on the histogram displaying the ICA frequency distribution. In line with Lemma 2, this is the ICA value that is obtained when all the unidentifiable correlations between the counterfactuals are assumed to equal zero, i.e., $\rho_{T_0T_1} = \rho_{T_1S_0} = \rho_{T_0S_1} = \rho_{S_0S_1} = 0$.

Finally, the association between both endpoints after adjusting for treatment and trial was estimated from model (5) as $\hat{R}_{ind} = 0.70$ (95% CI: [0.48; 0.83]). Notice that this value is obtained under the assumption $\rho_{T_0S_0} = \rho_{T_1S_1}$ and, therefore, it differs from the MEP-ICA.

7 Conclusions

A new definition of individual level surrogacy was given based on individual causal treatment effects and the ICA. This definition sets the individual level surrogacy in a more coherent context with respect to the trial level and bridges the two main paradigms for the evaluation of surrogate endpoints, i.e., the causal inference and meta-analytic paradigms. Indeed, in the former, individual causal treatment effects are often the primary building block for the analysis in a single-trial setting, whereas in the later expected causal treatment effects, i.e., the averages of the individual causal effects across all patients within the trial populations, are used to define trial level surrogacy. In the present work, causal inference methods were used to define individual level surrogacy in a meta-analytic context. In this way, both paradigms come together to create a more coherent methodology.

In addition, using the MEP, the underlying link between the new and classical definition of individual level surrogacy was established. It is important to point out that R_{ind} has scientific value in itself and, therefore, we do not envision the new proposal as a substitution of the classical definition of individual level surrogacy but rather as an addition to it.

In spite of all the assumptions made for the analysis of the ARMD data, a consistent pattern seems to emerge from Figure 1 indicating that visual acuity at 24 weeks may have some value as a surrogate endpoint for visual acuity 52 weeks. The results are, however, not conclusive. This lack of conclusiveness is not rare when dealing with complex problems like the evaluation of surrogate markers (Burzykowski *et al.*, 2005; Alonso *et al.*, 2017). Therefore, clinical and other qualitative arguments should always be juxtaposed with formal evaluation.

References

- 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.* , **71**, 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* (DOI: 10.1111/biom.12483).
- 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. (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.
- Burzykowski, T., Molenberghs, G. and Buyse, M. (2005). *The Evaluation of Surrogate Endpoints*. New York: Springer-Verlag.
- Bridle S. L., Hobson M. P. , Lasenby A. N., and Saunders R. A maximum-entropy method for reconstructing the projected mass distribution of gravitational lenses. *Monthly Notices of the Royal Astronomical Society*, 299–895.
- Buyse, M. and Molenberghs, G. (1998) The validation of surrogate endpoints in randomized experiments. *Biometrics* **54**, 1014–1029.
- 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.
- Conlon, A. S. C., Taylor, J. M. G., and Elliott, M. R. (2013). Surrogacy assessment using principal stratification when surrogate and outcome measures are multivariate normal. *Biostatistics* **14**, 1–18.
- Edwin T. J. (1982). On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, **70**(9), 939–952.
- Elliott, M.R., Li, Y., Taylor, J.M.G. (2013). Accommodating missingness when assessing surrogacy via principal stratification. *Clinical Trials* **10**, 363–377.

- Fleming T.R. and DeMets D.L. (1996). Surrogate endpoints in clinical trials: are we being misled? *Annals of Internal Medicine* **125**, 605–613
- Frangakis C.E. and Rubin D.B. (2002). Principal stratification in causal inference. *Biometrics*, **58**: 21–29.
- Gail, M.H., Pfeiffer, R., van Houwelingen, H.C., and Carroll, R.J. (2000). On meta-analytic assessment of surrogate outcomes. *Biostatistics* **1**, 231–246.
- Glunt W., Hayden T., Johnson C. R. , Tarazaga P. (1999) Positive definite completions and determinant maximization. *Linear algebra and its applications* **288**, 1–10.
- Grone R. , Johnson C. R. , Sa E. M., Wolkowicz H. (1984) Positive definite completions of partial hermitian matrices. *Linear algebra and its applications* **58** 109–124.
- Holland P.W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association* **81**, 945–960.
- Jaynes E. T. Where do we go from here? In C. R. Smith and J. W. T. Grandy, editors, *Maximum-Entropy and Bayesian Methods in Inverse Problems*, 21–58. Kluwer Academic Publishers.
- Jaynes E. T. *Probability theory: the logic of science*. Cambridge University Press.
- Kahan, B. C., Jairath, V., Dore, C. G., and Morris, T. P. (2014). The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies. *Trials* **15**: 139.
- MacKay D. J. C. Density networks and their application to protein modelling. In J. Skilling and S. Sibisi, editors, *Maximum Entropy and Bayesian Methods*, Cambridge, 259–268, Dordrecht, Kluwer.

- Montgomery D. Entropies for continua: Fluids and magnetofluids. In J. Skilling and S. Sibisi, editors, *Maximum Entropy and Bayesian Methods*, pages 304–314, Dordrecht, Kluwer.
- Pharmacological Therapy for Macular Degeneration Study Group (1997). Interferon α -IIA is ineffective for patients with choroidal neovascularization secondary to age-related macular degeneration. Results of a prospective randomized placebo-controlled clinical trial. *Archives of Ophthalmology* **115**, 865–872.
- Rosenbaum, P.R. and Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.
- Rubin, D. B. (1980). Randomization analysis of experimental-data the Fisher randomization test—Comment. *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.
- Shkedy Z., Barbosa F.T. (2005) Bayesian Evaluation of Surrogate Endpoints. In: Burzykowski T., Molenberghs G., Buyse M. (eds) *The Evaluation of Surrogate Endpoints. Statistics for Biology and Health*. Springer, New York, NY
- Van der Elst, W., Molenberghs, G., Alonso, A. (2016). Exploring the relationship between the causal-inference and meta-analytic paradigms for the evaluation of surrogate endpoints. *Statistics in Medicine*, **35**, 8, 1281–1298.

Figure 1: Top left: Expected causal treatment effects on the true endpoint expected causal treatment effects on the surrogate endpoint. The size of each point is proportional to the number of patients in the corresponding center. Top right: Frequency distribution of the ICA values compatible with the data. Bottom: True endpoint versus surrogate endpoint for all individual patients, after correction for treatment and center.