Research Article

Statistics
in Medicine

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

Assessing a surrogate predictive value: A causal inference approach.

Ariel Alonso[1,*], **Wim Van der Elst**[2] **& Paul Meyvisch**[3]

Several methods have been developed for the evaluation of surrogate endpoints within the causal-inference and meta-analytic paradigms. In both paradigms much effort has been made to assess the capacity of the surrogate to predict the causal treatment effect on the true endpoint. In the present work, the so-called surrogate predictive function (SPF) is introduced for that purpose, using potential outcomes. The relationship between the SPF and the individual causal association (ICA), a new metric of surrogacy recently proposed in the literature, is studied in detail. It is shown that the SPF, in conjunction with the ICA, can offer an appealing quantification of the surrogate predictive value. However, neither the distribution of the potential outcomes nor the SPF are identifiable from the data. These identifiability issues are tackled using a two-step procedure. In the first step, the region of the parametric space of the distribution of the potential outcomes, compatible with the data at hand, is geometrically characterized. Further, in a second step, a Monte Carlo approach is used to study the behavior of the SPF on the previous region. The method is illustrated using data from a clinical trial involving schizophrenic patients and a newly developed and user friendly R package *Surrogate* is provided to carry out the validation exercise. Copyright © 2015 John Wiley & Sons, Ltd.

Keywords: Surrogate endpoint, Causal inference, Sensitivity analysis, R package Surrogate

1. Introduction

Over the last decades, several strategies have been proposed for the evaluation of surrogate endpoints within the so-called causal-inference and meta-analytic paradigms [1, 2, 3]. In the former, individual causal treatment effects are often the primary building block for the analysis in a single-trial setting (STS), 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 carry out the validation exercise. In both paradigms attempts have been made to

[1] *I-BioStat, Katholieke Universiteit Leuven, B-3000 Leuven, Belgium.*
[2] *I-BioStat, Universiteit Hasselt, B-3590 Diepenbeek, Belgium.*
[3] *Janssen Pharmaceutica, Companies of Johnson & Johnson, Belgium.*
[*] *Correspondence to: Ariel Alonso, Kapucijnenvoer 35 blok d (bus 7001), 3000 Leuven, Belgium. Email: ariel.alonsoabad@kuleuven.be*

Contract/grant sponsor: The research leading to these results has received funding from the European Seventh Framework programme [FP7 2007 − 2013*] under grant agreement Nr.* 602552*.*

assess the capacity of the surrogate to predict the causal treatment effect on the true endpoint. Indeed, coefficients of determination and information-theoretic metrics have been introduced in the meta-analytic context, to assess the prediction of the expected causal treatment effect on the true endpoint using the expected causal treatment effect on the surrogate [1, 4, 5]. Similarly, in the causal-inference paradigm, the causal effect predictiveness (CEP) surface was proposed to evaluate the predictive value of a principal surrogate [6]. More recently, the so-called individual causal association (ICA), a metric with a direct interpretation in terms of prediction accuracy, has been introduced in a causal-inference framework as well [7, 8].

The validation of surrogate endpoints in the special setting in which one or both outcomes are binary, has already received attention in the literature [6, 9, 10]. For instance, a Bayesian modeling approach has been proposed to estimate the associative proportion when both the true and surrogate endpoints are binary, under the assumption of monotonicity [9]. The previous method was extended to accommodate missing data as well [10]. In the present work, the so-called surrogate predictive function is introduced to evaluate the surrogate predictive value when both endpoints are binary, using individual causal treatment effects in the STS. The methodology builds up on recently introduced validation strategies and allows to answer important scientific questions.

A common problem faced by many causal inference methods is their reliance on untestable assumptions to achieve identifiability of the parameters of interest. For example, to achieve identifiability when estimating the CEP, it has been assumed in previous research that the surrogate endpoint is constant in the control group, and that the value of the surrogate potential outcome for the new treatment (S_1) in the control group could be predicted using baseline covariates [6]. Assuming a constant value for the surrogate in the control group may be unrealistic in most practical situations and good predictive baseline covariates may actually not exist or may not be available. We address the identifiably issues using a two-step procedure. Basically, the surrogate predictive value is evaluated across different values of the unidentifiable parameters characterizing the distribution of the potential outcomes.

In section 2 the causal-inference model is presented. The SPF is introduced in section 3 where its relationship with the ICA is studied and the strategy to cope with the identifiability issues is described. The case study in presented and analyzed in section 4. In section 5 a simulation study is carried out to evaluate important aspects of the proposed methodology. The case study is re-analyzed in section 6 and some final comments are given in section 7.

2. Causal-inference model

In the rest of the manuscript it will be assumed that only two treatments are under evaluation ($Z = 0/1$) and both the true and surrogate endpoints are binary variables coded as 1 when a beneficial outcome is observed and 0 otherwise. In addition, the standard stable unit treatment value assumption (SUTVA) will also be made [11].

The so-called Rubin's model for causal inference assumes that each patient has a four dimensional vector of potential outcomes $\boldsymbol{Y} = (T_0, T_1, S_0, S_1)'$. 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 treatment or control, respectively. On account of simplicity, in the following the discussion will be temporarily restricted to the surrogate endpoint, but similar arguments can be put forward for the true endpoint as well.

The bivariate distribution of the vector of potential outcomes for the surrogate $\boldsymbol{Y}_S = (S_0, S_1)'$ is characterized by the parameters $\pi^S_{ij} = P(S_0 = i, S_1 = j)$ with $i, j = 0, 1$, and has marginals $\pi^S_{i\cdot} = \sum_j \pi^S_{ij}$, $\pi^S_{\cdot j} = \sum_i \pi^S_{ij}$. However, often in practice only one of the two potential outcomes S_0 and S_1 can be observed and, consequently, the distribution of $\boldsymbol{Y}_S$ is frequently not identifiable [12]. More specifically, the association structure of the two potential outcomes cannot be inferred from the data. Unlike the association structure, the marginal probabilities $\boldsymbol{\pi}_S = \left(\pi^S_{0\cdot}, \pi^S_{1\cdot}, \pi^S_{\cdot 0}, \pi^S_{\cdot 1}\right)'$ are identifiable under fairly general conditions. In fact, under SUTVA, $S = ZS_1 + (1 - Z)S_0$ and if the treatment assignment is independent of the potential outcomes ($\boldsymbol{Y}_S \perp Z$), then $\pi^S_{1\cdot} = E(S|Z = 0)$ with

$\pi^S_{0\cdot} = 1 - \pi^S_{1\cdot}$ and $\pi^S_{\cdot 1} = E\,(S|Z=1)$ with $\pi^S_{\cdot 0} = 1 - \pi^S_{\cdot 1}$. SUTVA basically states that the potential outcomes of an individual are independent of the treatments received by other individuals in the study and that the observed outcome under treatment Z equals the corresponding potential outcome S_Z. In addition, due to the random treatment allocation, the aforementioned assumption of independence $\boldsymbol{Y}_S \perp Z$ can often be guaranteed in randomized clinical trials.

As previously stated, in order to identify the entire bivariate distribution of $\boldsymbol{Y}_S$ additional assumptions on the association structure are needed. To this end, let us now consider the odds ratio $\theta_S = \pi^S_{00}\,\pi^S_{11}/\pi^S_{10}\,\pi^S_{01}$. Using θ_S and the marginal probabilities, the full bivariate distribution of $\boldsymbol{Y}_S$ can be recovered [13].

The individual causal effect of the treatment on the surrogate can be defined as $\Delta S = S_1 - S_0$; it follows a multinomial distribution parametrized by $\pi^{\Delta S}_i = P\,(\Delta S = i) = \sum_{pq} \pi^S_{pq}$ with $i = -1, 0, 1$ and the sum taken over all sub-indexes p, q satisfying $q - p = i$. Note that, like the distribution of $\boldsymbol{Y}_S$, the distribution of the individual causal treatment effect on the surrogate endpoint ΔS is not identifiable from the data, without making untestable assumptions about the association structure of the potential outcomes. Nonetheless, once θ_S is fixed, the distribution of ΔS becomes fully identifiable.

Similarly, the potential outcomes $\boldsymbol{Y}_T = (T_0, T_1)'$ can be used to define the individual causal treatment effect on the true endpoint ΔT 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 the fundamental quantity used in the following sections to assess the surrogate predictive value.

[Insert Table 1 about here]

3. Surrogate predictive value

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 [10]. Along these lines, it has been proposed to assess surrogacy using the so-called individual causal association (ICA), defined as the association between the individual causal treatment effects ΔT and ΔS [7, 8]. When both endpoints are continuous and normally distributed, the ICA can be quantified using the Pearson correlation coefficient $\rho_\Delta = \text{corr}\,(\Delta_T, \Delta_S)$ [7]. The previous quantification has been extended to binary endpoints using the following information-theoretic measure of association [8]

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

The term in the numerator is the so-called mutual information and it is defined as

$$I(\Delta T, \Delta S) = \sum_{i,j=-1}^{1} \pi^{\Delta}_{ij} \log\left(\frac{\pi^{\Delta}_{ij}}{\pi^{\Delta T}_i \pi^{\Delta S}_j}\right).$$

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. Furthermore, the denominator in (1) equals the minimum of the entropies of the individual causal treatment effects, which are defined as $H(\Delta T) = \sum_{i=-1}^{1} \pi^{\Delta T}_i \log(\pi^{\Delta T}_i)$, and $H(\Delta S) = \sum_{j=-1}^{1} \pi^{\Delta S}_j \log(\pi^{\Delta S}_j)$. The concept of entropy lies at the center of information theory and quantifies the randomness or uncertainty associated with a random variable [14, 15].

The ICA, as given in (1), 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. Indeed, $R^2_H(\Delta T, \Delta S)$ is invariant under one-to-one transformations and 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 $P\left[\Delta T = \psi(\Delta S)\right] = 1$ [8]. Consequently, when $R^2_H(\Delta T, \Delta S) = 1$ 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.

Even though $R^2_H(\Delta T, \Delta S)$ does provide a quantification of the surrogate predictive value, it does not give any information regarding the specific form of the prediction function ψ and leaves some important scientific questions unanswered. For instance, it does not allow to assess how likely it is that the treatment will have a negative impact on the true endpoint, given that it has a beneficial effect on the surrogate, i.e., the probability that the surrogate will produce a false positive result.

To explore these issues let us now consider a general prediction function $\psi : \{-1, 0, 1\} \to \{-1, 0, 1\}$. The predictive value of ψ can be assessed using the expression

$$\begin{aligned} P\left[\Delta T = \psi(\Delta S)\right] &= \sum_{i=-1}^{1} P\left(\Delta T = i, \psi(\Delta S) = i\right), \\ &= \sum_{i=-1}^{1} \sum_{j \in \psi^{-1}(i)} P\left(\Delta T = i, \Delta S = j\right), \\ &= \sum_{i=-1}^{1} \sum_{j \in \psi^{-1}(i)} P\left(\Delta T = i | \Delta S = j\right) P\left(\Delta S = j\right), \end{aligned} \tag{2}$$

where $\psi^{-1}(i) = \{j \in \{-1, 0, 1\} : \psi(j) = i\}$. The probabilities $P\left(\Delta S = j\right)$ do not involve the true endpoint and can be considered an intrinsic characteristic of the surrogate-endpoint-treatment pair.

On the other hand, the function $r : \{-1, 0, 1\}^2 \to [0, 1]$ given by $r(i, j) = P\left(\Delta T = i | \Delta S = j\right)$ fully captures the relationship between the individual causal treatment effects on the surrogate and true endpoint in (2). We shall denote r the surrogate predictive function (SPF), i.e., the function describing the full conditional distribution of ΔT given ΔS. The SPF allows to address some important scientific questions that cannot be explicitly answered only using $R^2_H(\Delta T, \Delta S)$. For instance, $r(-1, 1)$ quantifies the probability that the treatment has a negative impact on the true endpoint given that it has a beneficial impact on the surrogate, i.e., the probability that the surrogate will produce of a false positive result. Similarly, $r(1, -1)$ quantifies the probability that the treatment has a beneficial impact on the true endpoint given that it has a negative impact on the surrogate or, equivalently, the probability that the surrogate will produce a false negative result. It may be argued that $r(-1, 1) = r(1, -1) \approx 0$ is a desirable property for a good surrogate endpoint.

The SPF is also related to concepts previously introduced in the literature, for instance, it is intrinsically related to the concept of causal necessity proposed by Frangakis & Rubin [16]. These authors defined that S is necessary for the effect of treatment on the outcome T, if a causal effect of treatment on T can occur only if a causal effect of treatment on S has occurred. Essentially, causal necessity can be re-stated as $r(0, 0) = 1$. Another interesting conceptual setting is obtained when $P\left[\Delta T = \Delta S\right] = 1$, i.e., the treatment has identical individual causal effects on both endpoints. The following result fully characterizes the previous scenario using the SPF (the proof is straightforward).

Lemma 1 *Let T and S denote a binary true and surrogate endpoint respectively. Under the causal inference model introduced in Section 2, $P\left[\Delta T = \Delta S\right] = 1$ if and only if $r(i, j) = 0$ for all $i \neq j$.*

Furthermore, there is a close relationship between the SPF and the best prediction function associated with the distribution of $\mathbf{\Delta}$. To illustrate this let us first define the best prediction function as the function $\psi_b = \arg\max_\psi P\left[\Delta T = \psi(\Delta S)\right]$. The following lemma describes the relationship between the SPF and the best prediction function (Proof provided in the the Supplementary Materials accompanying the paper)

Lemma 2 *Let T and S denote a binary true and surrogate endpoint respectively. Further, let $\psi_b : \{-1, 0, 1\} \to \{-1, 0, 1\}$ be the function defined as*

$$\psi_b(j) = \arg\max_i r(i, j) = \arg\max_i P\left(\Delta T = i | \Delta S = j\right).$$

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. The function ψ_b is the best prediction function associated with the distribution of $\mathbf{\Delta}$.

Although methodologically appealing, the SPF is not identifiable from the data. In the following section a two-step procedure will be introduced to handle the identifiability issues.

3.1. Assessing the SPF

As previously stated, the SPF is not identifiable from the data and, consequently, cannot be directly estimated. In causal inference, this type of problem is often tackled by defining a number of identifiability assumptions. For instance, an assumption that is often used is the so-called monotonicity assumption. Under monotonicity $P(T_0 \leq T_1) = P(S_0 \leq S_1) = 1$ so, basically, $\pi_{10}^T = \pi_{10}^S = 0$. Identifiability conditions are frequently combined with additional modeling assumptions in order to estimate the parameters of interest. For instance, a Bayesian modeling approach has been used to estimate the associative (AP) and dissociative (DP) proportions (the definitions of AP and DP are given in the supplementary materials), where the unobserved potential outcomes were treated as missing data and imputation techniques were applied [9]. Identifiability was achieved under the assumption of monotonicity by selecting appropriate prior distributions for the unidentifiable parameters. A similar Bayesian approach to estimate the associative proportion under different monotonicity assumptions and missing data generating mechanisms has been proposed as well [10].

The use of identifiability conditions in this context 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. It is also important to point out that these issues are intrinsic to the use of potential outcomes and equally affect Bayesian and frequentist methods. Vansteelandt *et al.* [17], and some of the references there in, offer an in-depth discussion of the identifiability problem from a frequentist perspective.

Along the lines presented in Alonso *et al.* [8], we approach the identifiability problem following a two-step procedure, and based on the distribution of the vector of potential outcomes $\boldsymbol{Y}$. 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 $i, j, p, q = 0/1$. In a first step, we geometrically characterize the subspace $\Gamma_D \subset \Gamma$ compatible with the data at hand and, in a second step, study the behavior of the SPF on Γ_D. This approach is not aimed at estimating the true SPF, which is not identifiable, but it can better be thought of as a sensitivity analysis. In fact, each point in Γ_D can be conceptualized as a *world* compatible with ours and, therefore, the behavior of the SPF on Γ_D completely describes the surrogate predictive value across all scenarios compatible with the data.

In order to characterize Γ_D notice first that, as described in [9] and [10], the data at hand impose some restrictions on π_{ijpq}. Indeed, 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

$$\begin{aligned}
\pi_{1\cdot1\cdot} &= P(T=1,S=1|Z=0), \quad \pi_{\cdot1\cdot1} = P(T=1,S=1|Z=1),\\
\pi_{1\cdot0\cdot} &= P(T=1,S=0|Z=0), \quad \pi_{\cdot1\cdot0} = P(T=1,S=0|Z=1),\\
\pi_{0\cdot1\cdot} &= P(T=0,S=1|Z=0), \quad \pi_{\cdot0\cdot1} = P(T=0,S=1|Z=1),\\
\pi_{\cdot\cdot\cdot\cdot} &= 1,
\end{aligned} \tag{3}$$

with the points in the sub-indexes indicating sums over those specific sub-indexes. Further, if one defines the vector $\boldsymbol{b}' = (1, \pi_{1\cdot1\cdot}, \pi_{1\cdot0\cdot}, \pi_{\cdot1\cdot1}, \pi_{\cdot1\cdot0}, \pi_{0\cdot1\cdot}, \pi_{\cdot0\cdot1})$, then all the identified restrictions in (3) can be written as a system of linear equations,

$$\mathbf{A}\boldsymbol{\pi} = \boldsymbol{b}, \tag{4}$$

with $\mathbf{A}$ a binary matrix (details provided in the Supplementary Materials). The hyperplane (4) geometrically characterizes the subspace of Γ compatible with the data at hand, i.e., $\Gamma_D = \{\boldsymbol{\pi} \in \Gamma : \mathbf{A}\boldsymbol{\pi} = \boldsymbol{b}\}$.

In the second step, the behavior of the SPF on Γ_D (notice that SPF: $\Gamma_D \to [0,1]^6$) needs to be studied in order to evaluate the surrogate predictive value across all scenarios compatible with the data. In the terminology of Vansteelandt *et al.* [17], the values taken by the SPF on Γ_D can be considered an Honestly Estimated Ignorance Region (HEIR), because they express ignorance due to the no identifiability of $\boldsymbol{\pi}$.

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 SPF on $\boldsymbol{\pi}$ and the high dimensionality of the latter. We tackle 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. Essentially, points will be uniformly sampled on Γ_D and the SPF will be 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. Notice also that, in this conceptual framework, the sampling scheme should not be interpreted as a prior distribution quantifying the likelihood of the $\boldsymbol{\pi}$s but merely as a fair, unbiased procedure to select some of them to study the SPF. Similarly, the displayed histograms of the SPF should not be interpreted as a posterior probability distribution, but as a frequency distribution useful to visualize the behavior of the SPF on Γ_D.

To implement the sampling procedure notice that the binary matrix $\mathbf{A}$ in (4) has rank 7 and can be chosen so that $\mathbf{A} = (\mathbf{A}_r|\mathbf{A}_f)$ where $\mathbf{A}_r$ is a full column rank matrix and $\mathbf{A}_f$ denotes the submatrix given by the last 9 columns. Similarly, the vector $\boldsymbol{\pi}$ can be partitioned as $\boldsymbol{\pi}' = \left(\boldsymbol{\pi}_r'|\boldsymbol{\pi}_f'\right)$ with $\boldsymbol{\pi}_f$ the subvector given by the last 9 components of $\boldsymbol{\pi}$. Using these partitions (4) can be rewritten as $\mathbf{A}_r\boldsymbol{\pi}_r + \mathbf{A}_f\boldsymbol{\pi}_f = \boldsymbol{b}$ (details provided in in the Supplementary Materials). The following algorithm can then be used to sample points on Γ_D:

1. Select a grid of values $G = \{g_1,\ g_2,\ ...,\ g_K\}$ in $(0,1)$. The specific values of the grid have to be selected in order to guarantee numerical stability.
2. From $k = 1$ until K do
   (a) Using the *Randfixedsum* algorithm [18, 19] generate the 9 components of $\boldsymbol{\pi}_f$ uniformally in the hyperplane $\mathbf{1}'\boldsymbol{\pi}_f = g_i$.
   (b) Calculate $\boldsymbol{\pi}_r = \mathbf{A}_r^{-1}\left(\boldsymbol{b} - \mathbf{A}_f\boldsymbol{\pi}_f\right)$ and $\boldsymbol{\pi}' = \left(\boldsymbol{\pi}_r'|\boldsymbol{\pi}_f'\right)$.
   (c) Repeat steps 2a and 2b M times.
3. From these $K \times M$ $\boldsymbol{\pi}$ vectors select those with all components positive (the valid vectors $\boldsymbol{\pi} > \mathbf{0}$).

The distribution of the vector of individual causal effects $\mathbf{\Delta}$, given in Table 1, can then be obtained as

$$\begin{aligned}
\pi^{\Delta}_{-1-1} &= \pi_{1010}, & \pi^{\Delta}_{0-1} &= \pi_{0010} + \pi_{1110}, \\
\pi^{\Delta}_{1-1} &= \pi_{0110}, & \pi^{\Delta}_{-10} &= \pi_{1000} + \pi_{1011}, \\
\pi^{\Delta}_{-11} &= \pi_{1001}, & \pi^{\Delta}_{01} &= \pi_{0001} + \pi_{1101}, \\
\pi^{\Delta}_{11} &= \pi_{0101}, & \pi^{\Delta}_{10} &= \pi_{0100} + \pi_{0111}, \\
\pi^{\Delta}_{00} &= \pi_{0000} + \pi_{0011} + \pi_{1100} + \pi_{1111},
\end{aligned} \tag{5}$$

and $\pi^{\Delta}_{11} = 1 - \pi^{\Delta}_{-1-1} - \pi^{\Delta}_{0-1} - \pi^{\Delta}_{1-1} - \pi^{\Delta}_{-10} - \pi^{\Delta}_{00} - \pi^{\Delta}_{10} - \pi^{\Delta}_{-11} - \pi^{\Delta}_{01}$. Finally, based on these values, the SPF can be computed as $r(i,j) = \pi^{\Delta}_{ij}/\pi^{\Delta S}_{j}$ with $\pi^{\Delta S}_{j} = \sum_j \pi^{\Delta}_{ij}$. The results obtained from the previous sensitivity analysis can be summarized using the average surrogate predictive function $\bar{r}(i,j)$ defined as the average value of $r(i,j)$ across all values of $\boldsymbol{\pi}_{km}$.

4. Case study

A practical problem that is frequently encountered when validating surrogate endpoints is the lack of user-friendly software packages to conduct the analysis. The R package *Surrogate*, freely available at `https://cran.r-project.org/web/packages/Surrogate/index.html`, allows for the computation of the SPF and related metrics such as ICA. 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 study is provided and the implementation in R is discussed.

A clinical trial in Schizophrenia 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; [20]). 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; [21]) is a subscale of PANSS including only 18 items. 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 [22, 23].
Even though there is not a clear *gold* standard among psychiatric rating scales, in the present study PANSS is the most complete and reliable instrument and, therefore, it will be considered the main outcome or true endpoint. BPRS will be treated as the secondary outcome or potential surrogate endpoint. Basically, the main idea is to evaluate if a simpler and, hence, easier to administer scale like BPRS, could be reliably used as a substitute for the more complex PANSS scale that requires more time and expertise to be administered.

The individual causal association Four settings were considered in the analysis regarding monotonicity: i) Monotonicity holds for T, ii) Monotonicity holds for S, iii) Monotonicity holds for S and T, and iv) Monotonicity holds neither for S nor for T. Each of the previous settings defined a region in Γ_D where the Monte Carlo procedure was applied to study the behavior of R^2_H and the SPF. Due to its better numerical performance, a slight variation

of the algorithm introduced in Section 3.1 was used in the first step of the Monte Carlo procedure (details in Web appendix and *Surrogate* package manual) and a total of $M = 10000$ vectors $\boldsymbol{\pi}$ were sampled in each of the previous regions.

Table 2 summarizes the main findings and Figure 1 (second row) displays the frequency distributions of R^2_H for the different settings in the sensitivity analysis. The plots indicate that R^2_H tends to take larger values in the region where monotonicity does not hold than in the other regions of Γ_D. Actually, the monotonicity assumption has a major impact on the results, e.g., the mean R^2_H assuming no monotonicity is more than 4 times larger than the mean R^2_H assuming monotonicity for both S and T. Notice that all vectors $\boldsymbol{\pi}$ in Γ_D are compatible with the data at hand, i.e., it is impossible to discriminate between these regions based solely on the data. However, in some situations, domain-specific knowledge can be used to evaluate the plausibility of the different scenarios. Essentially, one wants to use biological knowledge to evaluate how likely it is that our reality lies in one specific region of Γ_D like, for instance, a region where monotonicity holds. This analysis can be carried out by using causal diagrams available in the package *Surrogate*. As a way of illustration let us consider the causal diagrams displayed in the first row of Figure 1, in these diagrams the two horizontal lines depict the identifiable informational coefficients of association between S and T in the two treatment conditions, i.e., $\widehat{r}^2_h(S_0,\, T_0) = 0.51$ and $\widehat{r}^2_h(S_1,\, T_1) = 0.60$. Essentially, these coefficients quantify the association between the surrogate and the true endpoint in both treatment groups and can be interpreted along the lines presented in [8].

The other four non-horizontal lines depict the medians of the unidentified informational coefficients of association between the counterfactuals. When monotonicity is not assumed (first causal diagram), the median informational association between the potential outcomes for the true and surrogate endpoints are small, i.e., $\widehat{r}^2_h(S_0,\, S_1) = \widehat{r}^2_h(T_0,\, T_1) = 0.10$. This suggests that a patient's outcome on BPRS/PANSS in the active control condition (S_0/T_0) conveys little information on the patient's outcome on BPRS/PANSS in the experimental treatment condition (S_1/T_1). Given that the treatments under study are similar and S_0, S_1 (and also T_0, T_1) are repeated measurements on the same patient, this weak association may be considered counter-intuitive. Further, the other median informational associations $\widehat{r}^2_h(S_0,\, T_1) = 0.11$ and $\widehat{r}^2_h(S_1,\, T_0) = 0.09$ are also low. Since the BPRS is a sub-scale of the more complex PANSS scale, one would also expect a certain level of association between these potential outcomes and independence is again counter-intuitive.

When monotonicity is assumed for S alone (second causal diagram), the median informational associations between the potential outcomes are substantially larger. For example, the median $\widehat{r}^2_h(S_0,\, S_1) = \widehat{r}^2_h(T_0,\, T_1) = 0.67$, $\widehat{r}^2_h(S_0,\, T_1) = 0.50$ and $\widehat{r}^2_h(S_1,\, T_0) = 0.46$. As stated in the previous paragraph, this pattern of associations between the potential outcomes seems to be more compatible with our biological expectations. Although this assessment is only meant for illustrative purposes, a similar analysis can be done to bring expert opinion into the evaluation process (details in the Supplementary Materials).

[Insert Figure 1 about here]

In general, the individual causal treatment effect on BRRS does not seem to convey a lot of information on the individual causal treatment effect on PANSS. However, one may still wonder if, for example, a lack of treatment effect on BPRS ($\Delta S = 0$) may be indicative of a lack of treatment effect on PANSS as well ($\Delta T = 0$). The SPF allows to zoom in and analyze the prediction problem in more detail.

[Insert Table 2 about here]

The Surrogate Predictive Function Figures 2 and 3 summarize the behavior of the SPF under the no monotonicity and monotonicity for S regions, using histograms. Due to space constraints, the SPF histograms that

were obtained under the monotonicity for T and monotonicity for both S and T scenarios are not given here, but they are provided in the Supplementary Materials. In general the results were similar to those presented below.

[Insert Figures 2 and 3 about here]

As shown in Figure 2 (bottom left figure), in the no monotonicity region, $r(-1,1) = P(\Delta T = -1|\Delta S = 1)$ does not seem to take values larger than 0.25 (mean of $r(-1,1) = 0.0443$, max = 0.2324) and, therefore, the probability of a false positive result seems to be rather small. There is also some evidence that the probability of a false negative result (top right figure) may be small as well, mean $r(1,-1) = P(\Delta T = 1|\Delta S = -1) = 0.0483$, but now the range for this probability is much wider $[0.0001; 0.6067]$, hinting on a substantial level of uncertainty due to the unidentifiability of this parameter. It is clear from the analysis that a negative effect on BPRS seems to mostly lead to a negative effect on PANSS, but there is a rather large level of uncertainty as the histograms of $r(-1,-1) = P(\Delta T = -1|\Delta S = -1)$ and $r(0,-1) = P(\Delta T = 0|\Delta S = -1)$ clearly show.
The results displayed in the center figure offer some degree of support for causal necessity (as defined by [16]), with mean $r(0,0) = P(\Delta T = 0|\Delta S = 0) = 0.8597$. Thus, a lack of effect on BPRS seems to give evidence of a lack of effect on PANSS. There is still some degree of uncertainty in this case with $r(0,0)$ taking values between 0.5476 and 0.9705, however, it is smaller than the one observed when the treatment had a negative impact on the surrogate. Similar results are obtained when the treatment has a beneficial effect on BPRS (last row of the figure). Overall, there is some evidence that $r(i,i) = P(\Delta T = i|\Delta S = i)$ may be large for all i (the main diagonal in the figure), with all means ≥ 0.7484, all medians ≥ 0.8251 and all modes ≥ 0.8711. However, as the previous discussion indicates, the lack of indentifiability introduces a substantial level of uncertainty regarding the true value of these probabilities.
The results obtained in the scenario where monotonicity holds for S were very interesting. Notice that no results are shown for $r(i, j = -1)$ in this setting, as the probabilities of these events are 0 when monotonicity is assumed for S. As it can be seen in Figure 3 (top center figure), the mean $r(0,0) = 0.9091$ (max = 0.9715) and, therefore, causal necessity seems to be largely supported, i.e., a lack of treatment effect on BPRS seems to give strong evidence of a lack of treatment effect on PANSS. However, when there is a positive individual causal treatment effect on S ($\Delta S = 1$), there is substantial uncertainty with respect to the individual causal treatment effect on T (ΔT) (see Figure 3, second row). Indeed, the mean $r(1,1) = 0.5153$ (max = 0.9247), the mean $r(-1,1) = 0.1213$ (max = 0.4446) and mean $r(0,1) = 0.3633$ (max = 0.7667). Thus, although a negative effect on PANSS does not seem to be very likely when a positive effect on BPRS is observed, there is still a very large amount of uncertainty regarding the individual causal treatment effect on PANSS in this setting, as the histograms of $r(0,1)$ and $r(1,1)$ clearly indicate.
The previous analyses illustrate that the ICA and the SPF provide complementary and useful information for the validation exercise. In fact, in both scenarios the ICA indicated that accurate predictions of ΔT using ΔS were not generally possible. However, the ICA only offers a global quantification of the surrogate predictive value. In addition to that, the SPF offers a more detailed view of the different predictions scenarios and permits to identify the concrete situations in which the predictions may still be reasonably accurate and those situations in which they completely fail.

5. Impact of ignoring sampling variability: A simulation study

In Section 3.1, Γ_D was defined using the estimated components of $\boldsymbol{b}$ and the sampling variability of these estimates 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. In the present section, a simulation study is carried out to evaluate this issue, i.e.,

the impact of using $\widehat{\mathbf{b}}$ instead of $\mathbf{b}$, on the assessment of the ICA and SPF.

5.1. Simulation design

Table 3 shows the two scenarios considered for the identifiable marginal probabilities contained in $\mathbf{b}$. Notice that, in both scenarios, the surrogate and true endpoint are associated in the control and treated groups. Actually, in practice, the presence of an association between the putative surrogate and the true endpoint is often taken as a prerequisite for surrogacy and, therefore, we did not consider settings in which both endpoints were independent. In scenario 1 both endpoints are moderately associated with $\theta_{ST|Z} = 2.25$. Scenario 2 represents the more extreme and probably more unrealistic setting in which both endpoints are almost deterministically related in both treatment groups, with $P(T = S|Z) = 0.9$ and $\theta_{ST|Z} = 81$. Notice that, even though it may be unlikely in practice, scenario 2 is still methodologically and conceptually interesting. Notice also that, given that BPRS is a subscale of PANSS, in the case study the association between the surrogate and true endpoint is similar to the one considered in scenario 2. Furthermore, five sample sizes were evaluated, namely, $N = 50,\ 100,\ 300,\ 600$ and 1000 patients. For each sample-size-scenario combination, 250 data sets were generated using draws from a multinomial distribution. Thus, in total, $2,500$ data sets were obtained and in each of these data sets $\widehat{\mathbf{b}}$ was determined.

[Insert Table 3 about here]

Finally, the ICA and SPF were assessed using $\mathbf{b}$ ($\text{ICA}_{\mathbf{b}}$, $\text{SPF}_{\mathbf{b}}$), i.e., the true values given in Table 3, and the estimated values $\widehat{\mathbf{b}}$ ($\text{ICA}_{\widehat{\mathbf{b}}}$, $\text{SPF}_{\widehat{\mathbf{b}}}$) as the input for the proposed algorithm (more details in Web Appendix). The Monte Carlo procedure was implemented using $M = 50,000$ runs and assuming no monotonicity.

Main outcomes of interest: The main goal of the simulation study was the assessment of the bias induced by replacing $\mathbf{b}$ by $\widehat{\mathbf{b}}$ when analyzing the data. Therefore, the relative ICA bias, computed as $E\left[\left(\text{ICA}_{\widehat{\mathbf{b}}} - \text{ICA}_{\mathbf{b}}\right) / \text{ICA}_{\mathbf{b}}\right]$ was one of the studied outcomes. A similar outcome was also considered for the SPF.

5.2. Simulation results

Table 4 displays the results obtained in both scenarios. With respect to the ICA, the results showed that the bias induced by ignoring the sample variability is mostly negligible. Only when the sample size was rather small, i.e., $N = 50$ patients, a certain degree of bias was observed, but it never exceeded 15%. Importantly, for a sample size smaller than the one of the case study, i.e., $N = 300$, the relative bias was only about 1.3% in both scenarios.
With respect to the SPF, the relative bias in scenario 1 was always less than about 4% for samples of size $N = 100$ and always less than 2% for samples of size $N = 300$ or larger. Interestingly, in scenario 2, although the relative bias was generally small, large relatively biases were observed for $r\,(-1,\ 1)$ and $r\,(1,\ -1)$. For example, when $N = 300$, the relative bias for these parameters was about 11%. As expected, for sample sizes larger than $N = 600$, the relative bias was much smaller. Actually, for $N = 1000$ the relative bias was always smaller than 6%.
Summarizing, the previous results suggest that ignoring the sampling variability in $\widehat{\mathbf{b}}$ induces a negligible bias in the assessment of the ICA for sample sizes of $N = 100$ patients or larger. Additionally, the relative bias observed when assessing the SPF could be considered generally acceptable for moderate sample sizes ($N \geq 300$), taking values smaller than about 11% in both scenarios. However, there were some substantial differences in the relative bias for the SPF in scenarios 1 and 2, and more simulations may be needed to examine this issue in more detail.

[Insert Table 4 about here]

6. Accounting for the sampling variability

In the analyses presented in Section 4, the sampling variability in the estimates of the marginal probabilities contained in $\boldsymbol{b}$ was not taken into account. For example, $\pi_{1\cdot 1\cdot}$ was fixed at its estimated value 0.4215 in each run of the algorithm. In the previous section it was shown, via simulations, that ignoring sampling variability may produce a relative bias as large as 11% in the estimation of some components of the SPF. To account for the uncertainty in the estimation of $\pi_{1\cdot 1\cdot}$, this parameter can be uniformly sampled from its corresponding confidence interval $CI_{95\%} = [0.3562;\ 0.4868]$ at each run of the Monte Carlo algorithm and a similar procedure can also be used for the other marginal probabilities. Basically, this new version of the algorithm could be seen a second level sensitivity analysis that takes into account, on one hand, the uncertainty emanating from the unidentifiability of the distribution of the potential outcomes and, on the other hand, the uncertainty emanating from the estimation of its identifiable marginal probabilities.

In the following the case study is re-analyzed using the revised version of the algorithm that takes the sampling variability into account. In order to keep the length of the manuscript at a reasonable level, we will only present the results for the settings where monotonicity holds neither for S nor for T (no monotonicity) and the setting where monotonicity holds only for S. A throughout re-analysis of the case study and the corresponding implementation in R can be found in the Web Appendix.

Overall, both analyses produced similar results in the region where monotonicity does not hold (see Table 5). In addition, in the monotonicity for S region (see Table 6), $r\,(0,0)$ was rather high (minimum value larger than 0.77 irrespectively of the analysis) and $r\,(1,0)$, $r\,(-1,0)$ were rather low (maximum value smaller than 0.089 and 0.15 respectively) irrespectively of the analysis that was carried out. Thus, a lack of effect on BPRS ($\Delta S = 0$) seems to be indicative of a lack of effect on PANSS ($\Delta T = 0$), whether the sampling variability is taken into account or not. However, the assessment of $r\,(i, j = 1)$ differed substantially in both analysis. In general, when the sampling variability was accounted for, the measures of central tendency were larger for $r\,(1,1)$ and smaller for $r\,(-1,1)$, $r\,(0,1)$.

Interestingly, although accounting for the sampling variability, as might be expected, often led to an increased range (maximum-minimum) for the estimands of interest, sometimes it also led to a decrease in this range. Clearly, further simulation studies and theoretical developments will be needed to fully understand the complex interplay between all the involved factors like unidentifibility uncertainty and sampling uncertainty, among others. If the results of both analyses, with and without accounting for sampling variability, coincide, then one may feel more confident about the validity of the conclusions. However, when both methods lead to discordant results more caution is needed. Importantly, as one would theoretically expect, the simulation study showed that for large sample sizes both procedures deliver similar outputs and substantial differences mostly appear when the sample size is small. One may wonder, however, whether undertaking a complex task like the validation of a surrogate endpoint, is meaningful when only information on one, small clinical trial is available.

[Insert Tables 5 and 6 about here]

7. Conclusions

In the preceding sections it has been shown that the SPF can add important information to the analysis based on the ICA, a global measure of prediction accuracy recently introduced in the literature. The methodology proposed has, however, some practical and conceptual limitations. For instance, at the conceptual level, it should be pointed out that the proposed methodology cannot be easily classified into one of the main inferential frameworks, namely, the frequentist, Bayesian and likelihood inferential schools. This is an important conceptual issue that affects many other statistical procedures as well. For instance, Empirical Bayes methods are, strictly speaking, neither Bayesian nor frequentists (they obviously do not belong to the likelihood school neither). Similarly, methods like regression and classification trees, random forests and many other data mining algorithms cannot be classified into one of the three main inferential frameworks, in spite of being useful prediction tools. In the manuscript we follow a pragmatic approach and argue that a methodology that cannot be easily classified into one of the main inferential frameworks, may still be an approximate, useful solution to a relevant problem.

At the practical level our simulation studies indicate that ignoring the sampling variability in the estimate of $\boldsymbol{b}$ may produce moderate relative biases when assessing the SPF. A possible solution to this problem is to uniformly sample the components of $\boldsymbol{b}$ from their corresponding confidence intervals at each run of the Monte Carlo algorithm. This strategy is also implemented in the *Surrogate* package and in the Web appendix a detailed analysis of the case study is provided using this correction.

Missing data permeate medical research. The Monte Carlo approach introduced in Section 3.1 uses the data only to compute the sufficient statistics given in the components of $\boldsymbol{b}$. In the presence of missing data, methods like, for example, multiple imputation could be applied in a straightforward fashion to estimate these probabilities. In such a situation, the reliability of the results will be limited by the validity of the assumptions made to handle the missing observations. For instance, if classical multiple imputation is employed, then the results of the evaluation exercise will be valid only under the assumption of missing at random.

Supplementary Material

A Web Appendix that details the analyses of the case study (using the R package *Surrogate* and that provides a proof for Lemma 2 is available at the website of Statistics in Medicine.

References

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

[2]. Taylor JMG, Wang Y, Thiébaut, R. Counterfactual links to the proportion of treatment effect explained by a surrogate marker. *Biometrics* 2005; **61**:1102–1111.

[3]. VanderWeele T. Surrogate measures and consistent surrogates. *Biometrics* 2013; **69**: 561–565.

[4]. Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H. The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics* 2000; **1**: 49–67.

[5]. Alonso A, Molenberghs G. Surrogate marker evaluation from an information theoretic perspective. *Biometrics* 2007; **63**: 180–186.

[6]. Gilbert PB, Hudgens MG. Evaluating candidate principal surrogate endpoints. *Biometrics* 2008; **64**: 1146–1154.

[7]. Alonso A, Van der Elst W, Molenberghs G, Buyse M, Burzykowski T. On the relationship between the causal-inference and meta-analytic paradigms for the validation of surrogate endpoints. *Biometrics* 2015; **71**: 15–24.

[8]. Alonso A, Van der Elst W, Molenberghs G, Buyse M, Burzykowski T. A causal-inference approach for the validation of surrogate endpoints based on information theory and sensitivity analysis. *Biometrics* 2016, Accepted.

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

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

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

[12]. Holland PW. Statistics and Causal Inference. *Journal of the American Statistical Association* 1986; **81**: 945–960.

[13]. Plackett RL. A class of bivariate distributions. *Journal of the American Statistical Association* 1965; **60**: 516–522.

[14]. Joe H. Relative entropy measures of multivariate dependence. *Journal of the American Statistical Association* 1989; **84**: 157–164.

[15]. Cover T, Tomas J. *Elements of Information Theory.* New York: Wiley, 1991.

[16]. Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics* 2002; **58**: 21–29.

[17]. Vansteelandt, S., Goetghebeur, E., Kenward, M. G. and Molenberghs, G. (2006). Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statist. Sinica* **16**, 953–979.

[18]. Emberson P, Stafford R, Davis R. Techniques for the synthesis of multiprocessor tasksets. Paper presented at 1st International Workshop on Analysis Tools and Methodologies for Embedded and Real-time Systems, Brussels, July 6th 2010, 6–11.

[19]. Stafford R. Random vectors with fixed sum [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/9700

[20]. Singh M, Kay S. 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* 1975; **43**: 103–113.

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

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

[23]. 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.

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

Table 2. Descriptives of ICA under different monotonicity scenarios.

Monotonicity scenario	Mean	Median	Mode	SD	Range
No monotonicity	0.5280	0.5475	0.5654	0.0964	[0.2352; 0.6951]
Monotonicity for T	0.2411	0.2439	0.2585	0.1309	[0.0034; 0.5599]
Monotonicity for S	0.2695	0.2633	0.2718	0.1383	[0.0219; 0.6114]
Monotonicity for S and T	0.1304	0.0858	0.0131	0.1348	[0.0001; 0.6322]

Table 3. Different scenarios (true marginal probabilities) that were used to simulate the data.

Scenario 1								Scenario 2							
$Z=0$				$Z=1$				$Z=0$				$Z=1$			
		T				T				T				T	
		0	1			0	1			0	1			0	1
S	0	0.30	0.20	S	0	0.30	0.20	S	0	0.45	0.05	S	0	0.45	0.05
	1	0.20	0.30		1	0.20	0.30		1	0.05	0.45		1	0.05	0.45

Table 4. Scenario 1: Relative bias in the estimation of R^2_H and $r\,(i,\ j)$ as a function of N.

Scenario 1					
			N		
Parameter	50	100	300	600	1000
R^2_H	0.0855	0.0352	0.0137	0.0044	−0.0032
$r\,(1,\ 1)$	0.0254	0.0204	0.0063	0.0027	−0.0011
$r\,(-1,\ 1)$	−0.0677	−0.0410	−0.0127	0.0016	0.0089
$r\,(0,\ 1)$	0.0125	0.0061	0.0054	0.0025	0.0033
$r\,(1,\ 0)$	0.0482	0.0302	0.0084	−0.0017	0.0080
$r\,(-1,\ 0)$	0.0038	−0.0002	0.0140	0.0031	−0.0050
$r\,(0,\ 0)$	−0.0208	−0.0111	−0.0072	0.0014	0.0010
$r\,(1,\ -1)$	0.0076	0.0153	0.0126	0.0054	0.0023
$r\,(-1,\ -1)$	−0.0419	−0.0337	−0.0196	−0.0055	−0.0012
$r\,(0,\ -1)$	0.0356	0.0251	0.0160	0.0072	0.0055
Scenario 2					
			N		
Parameter	50	100	300	600	1000
R^2_H	−0.1479	−0.0239	0.0130	0.0069	0.0047
$r\,(1,\ 1)$	0.0297	0.0193	0.0116	0.0047	0.0011
$r\,(-1,\ 1)$	−0.5202	−0.3199	−0.1138	−0.0601	−0.0249
$r\,(0,\ 1)$	0.0321	0.0208	−0.0115	0.0030	0.0075
$r\,(1,\ 0)$	0.0882	0.0384	−0.0154	−0.0199	−0.0198
$r\,(-1,\ 0)$	0.0026	−0.0149	−0.0123	−0.0167	−0.0046
$r\,(0,\ 0)$	−0.0086	−0.0016	0.0033	0.0042	0.0031
$r\,(1,\ -1)$	−0.3371	−0.2254	−0.1095	−0.0743	−0.0566
$r\,(-1,\ -1)$	−0.0043	0.0011	0.0054	0.0061	0.0070
$r\,(0,\ -1)$	0.1347	0.0726	0.0156	−0.0004	−0.0093

Table 5. SPF summary statistics under the no monotonicity assumption when the sampling variability in the marginal probabilities is not accounted for (left) and is accounted for (right)

	Sampling variability marginals not accounted for					Sampling variability marginals accounted for				
	Mean	Median	Mode	*SD*	[min; max]	Mean	Median	Mode	*SD*	[min; max]
$r(-1,\,-1)$	0.7484	0.8251	0.8711	0.2172	[0.0585; 0.9717]	0.8050	0.8359	0.8648	0.1347	[0.0087; 0.9930]
$r(0,\,-1)$	0.2033	0.1357	0.1034	0.1846	[0.0047; 0.7679]	0.1748	0.1463	0.1138	0.1272	[0.0033; 0.9232]
$r(1,\,-1)$	0.0483	0.0235	0.0150	0.0832	[0.0001; 0.6067]	0.0202	0.0161	0.0063	0.0190	[0.0001; 0.2134]
$r(-1,\,0)$	0.0781	0.0582	0.0421	0.0561	[0.0060; 0.2521]	0.1077	0.0907	0.0545	0.0753	[0.0008; 0.7185]
$r(0,\,0)$	0.8597	0.8906	0.9153	0.0874	[0.5476; 0.9705]	0.8043	0.8357	0.8733	0.1155	[0.0599; 0.9930]
$r(1,\,0)$	0.0622	0.0473	0.0395	0.0481	[0.0037; 0.3107]	0.0880	0.0736	0.0586	0.0606	[0.0023; 0.5438]
$r(-1,\,1)$	0.0443	0.0310	0.0224	0.0454	[0.0007; 0.2324]	0.0312	0.0274	0.0093	0.0230	[0.0001; 0.1593]
$r(0,\,1)$	0.1303	0.1118	0.0790	0.0896	[0.0044; 0.4287]	0.1261	0.1108	0.0920	0.0673	[0.0039; 0.5064]
$r(1,\,1)$	0.8254	0.8473	0.8838	0.0998	[0.4556; 0.9740]	0.8472	0.8585	0.8878	0.0728	[0.4715; 0.9939]

Table 6. SPF summary statistics under the monotonicity for S assumption when the sampling variability in the marginal probabilities is not accounted for (left) and is accounted for (right)

	Sampling variability marginals not accounted for					Sampling variability marginals accounted for				
	Mean	Median	Mode	*SD*	[min; max]	Mean	Median	Mode	*SD*	[min; max]
$r(-1,\ 0)$	0.0281	0.0249	0.0173	0.0167	[0.0028; 0.0670]	0.0768	0.0775	0.0865	0.0347	[0.0249; 0.1486]
$r(0,\ 0)$	0.9091	0.9074	0.9045	0.0246	[0.8456; 0.9715]	0.8774	0.8832	0.8928	0.0455	[0.7707; 0.9361]
$r(1,\ 0)$	0.0629	0.0612	0.0589	0.0151	[0.0201; 0.0899]	0.0459	0.0444	0.0402	0.0177	[0.0139; 0.0807]
$r(-1,\ 1)$	0.1213	0.1172	0.0512	0.0907	[0.0005; 0.4446]	0.0357	0.0333	0.0360	0.0288	[0.0005; 0.1051]
$r(0,\ 1)$	0.3633	0.3408	0.3255	0.1789	[0.0520; 0.7667]	0.1444	0.1442	0.1039	0.0702	[0.0326; 0.2898]
$r(1,\ 1)$	0.5153	0.50723	0.4783	0.1838	[0.1018; 0.9247]	0.8199	0.8030	0.7632	0.0726	[0.7034; 0.9541]

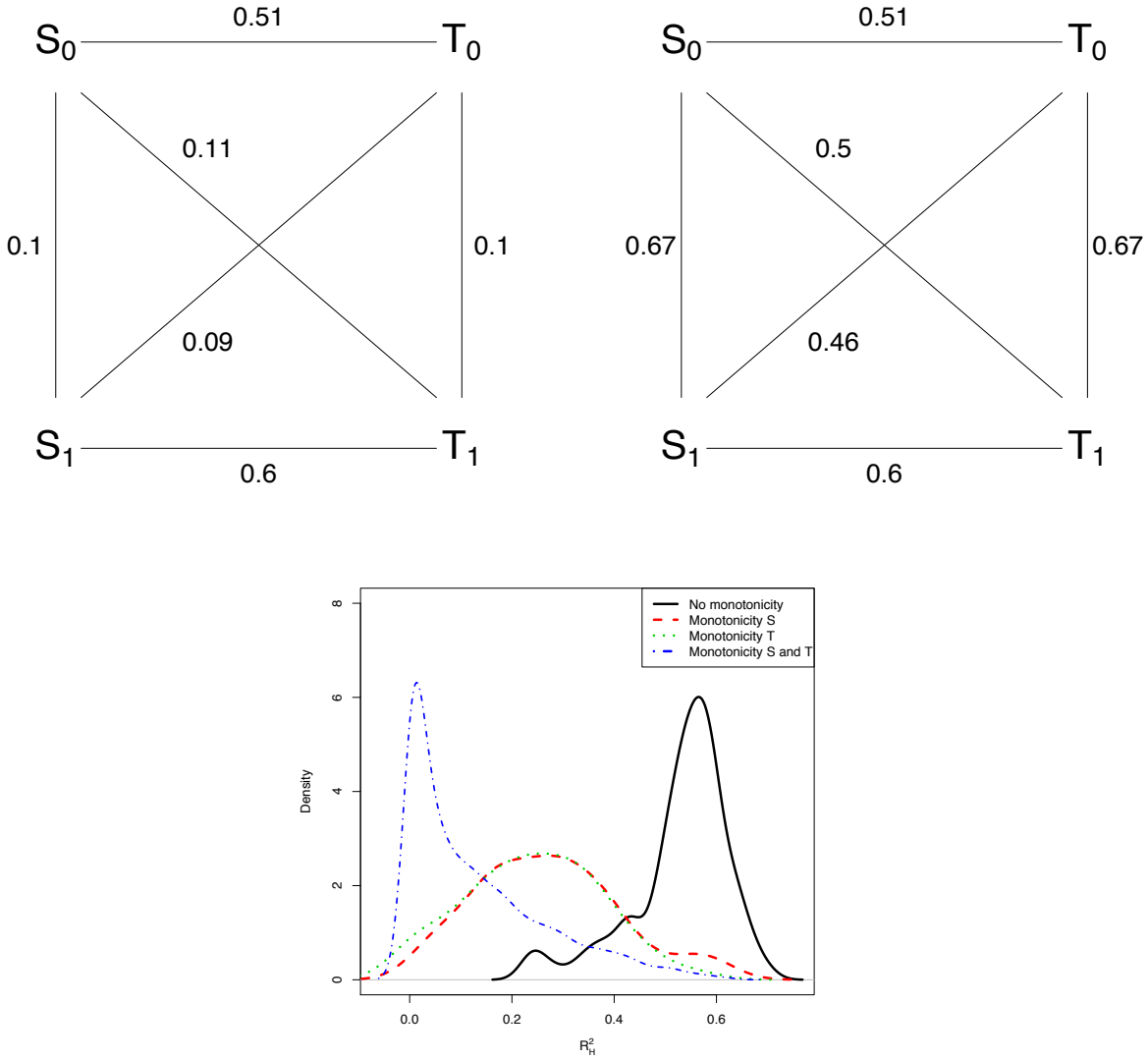


Figure 1. Causal diagrams in the no monotonicity (top left) and monotonicity for S (top right) scenarios. Frequency distribution of the ICA under the different monotonicity assumptions (second row).

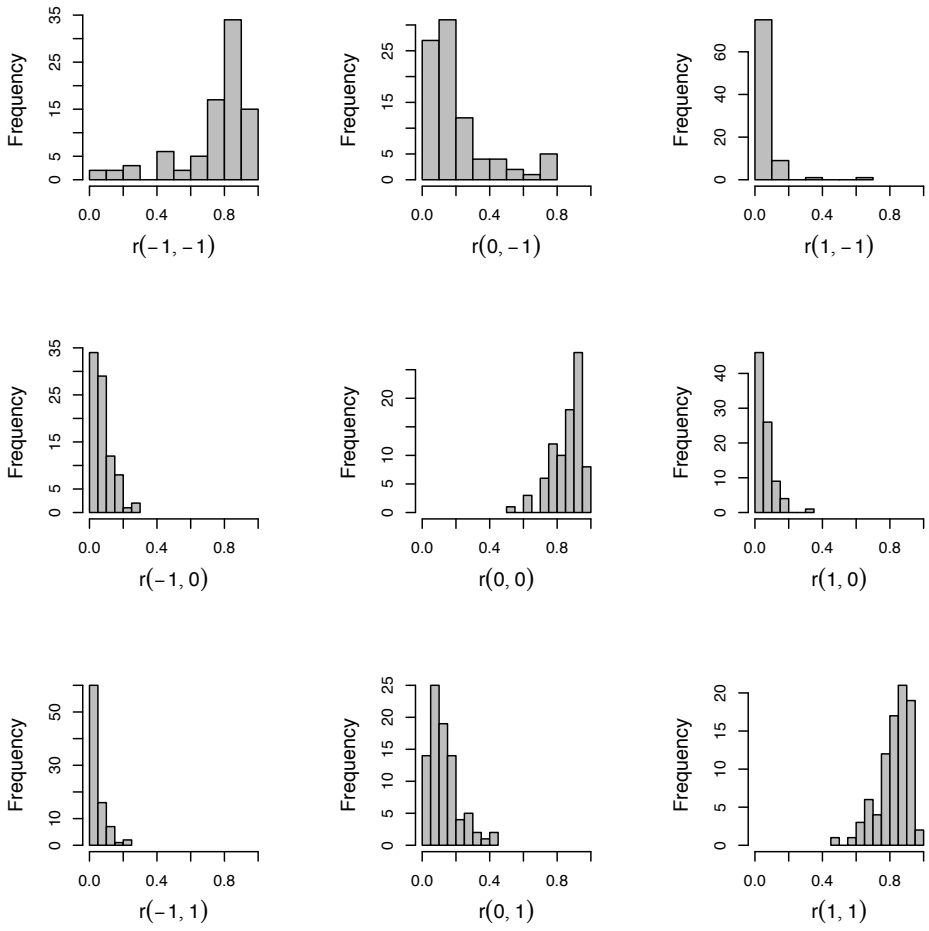


Figure 2. SPF, i.e., $r(i,j) = P\left(\Delta T = i | \Delta S = j\right)$ assuming no monotonicity in the case study.

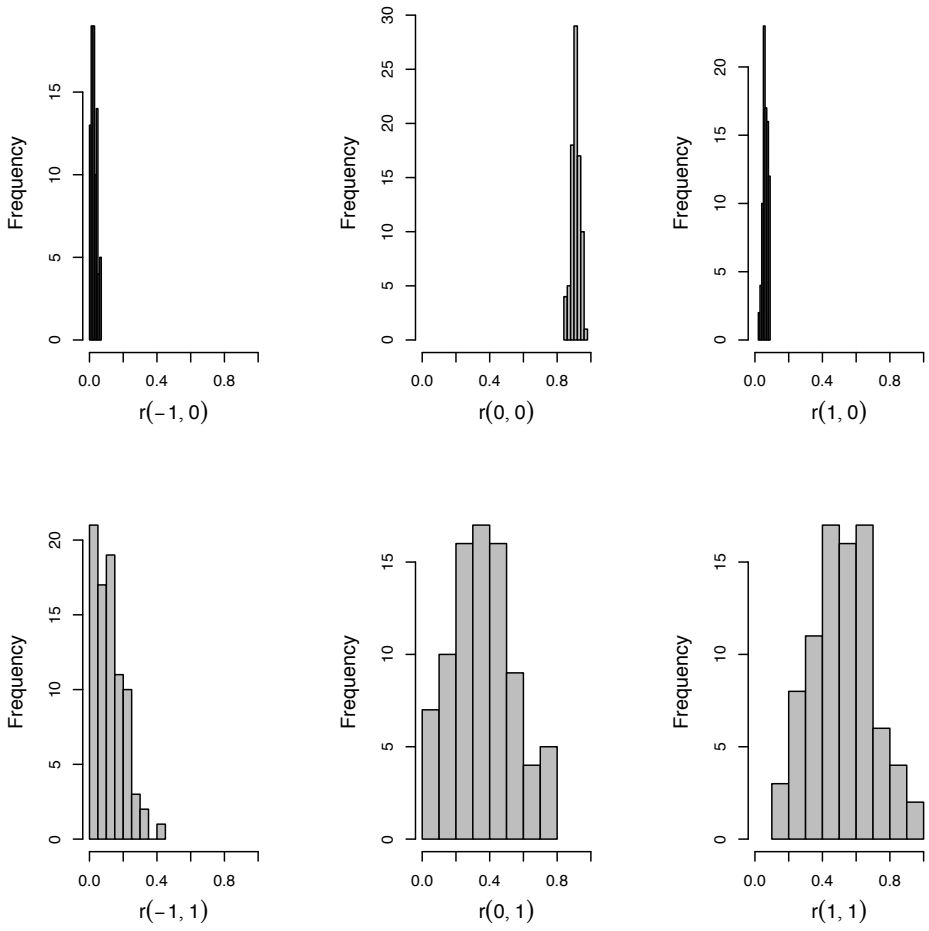


Figure 3. SPF, i.e., $r(i,j) = P\left(\Delta T = i | \Delta S = j\right)$ assuming monotonicity for S in the case study.

Web Appendix to 'Assessing a surrogate predictive value: A causal inference approach.'

Ariel Alonso, Wim Van der Elst & Paul Meyvisch

1 Analysis of the case study

This Appendix illustrates the use of the R package *Surrogate*, available at CRAN (`http://cran.r-project.org/web/packages/Surrogate/`), for the analysis of the case study described in Alonso *et al.* (2016b).

The data of the case study are introduced in Section 2 and analyzed in Section 3. The focus of the analysis will be on two metrics. First, the individual causal association (ICA; Alonso *et al.*, 2016a), that offers a general quantification of the surrogate predictive value in a single index. Second, the surrogate predictive function (SPF; Alonso *et al.*, 2016b), which allows for a more fine-grained assessment of how the individual causal effect on S can predict the individual causal effect on T. In Section 3, the data are analyzed ignoring the sampling variability in the estimates of the marginal probabilites used to deifine Γ_D. In Section 5, this sampling variability is taken into account in the analyses. The impact of ignoring the sampling variability on the results obtained in the sensitivity analysis is further study in Section 4 via simulations. Finally, in Section 6 some algebraic results and definitions are presented.

2 The dataset: a clinical trial in schizophrenia

The data come from a clinical trial designed to compare the efficacy of risperidone (experimental treatment) and haloperidol (control treatment) 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. In psychiatry several measures can be considered to assess a patient's global condition. A useful and sufficiently sensitive assessment scale is the Positive and Negative Syndrome Scale (PANSS; Singh & Kay (1975)). 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; Overall & Gorham (1962)) is a subscale of PANSS including only 18 items. 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 posttreatment scores relative to baseline scores (Kane *et al.*, 1988; Leucht *et al.*, 2005).

The dataset (`Schizo_Bin`) is included in the *Surrogate* package. After installation of the package in R, the following code can be used to load the package and the schizophrenia dataset in memory for the subsequent analyses:

```
library(Surrogate) # load the Surrogate library
data(Schizo_Bin) # load the data
head(Schizo_Bin) # have a look at the first observations

##   Id InvestId BPRS_Bin PANSS_Bin Treat
## 1  2      104        1         1     1
```

```
## 2  6      104        0         0    -1
## 3  7      104        0         0     1
## 4  8      104        0         0    -1
## 5 14       26        1         0    -1
## 6 18       26        1         1     1
```

The dataset contains five variables:

- 'Id': the identification number of the patient.
- 'InvestId': the identification number of the treating physician.
- 'Treat': the treatment indicator. -1 = control treatment (dose of 10 mg. of haloperidol), 1 = experimental treatment (dose of 8 mg. of risperidone).
- 'BPRS_Bin': a binary endpoint taking values: 1 = clinically meaningful change has occurred, 0 = otherwise.
- 'PANSS_Bin': a binary endpoint taking values: 1 = clinically meaningful change has occurred, 0 = otherwise.

In the analyses below, it will be examined whether clinically meaningful change on BPRS, a simpler and easier to administer scale, is an appropriate surrogate for clinically meaningful change on PANSS, a more complex scale that requires more time and more skilled personnel for its administration. To simplify the exposition, in the following sections the names BPRS and PANSS will be loosely used to refer to clinically meaningful change in these scales.

3 Analysis of the case study: Implementation in R

3.1 Exploratory data analysis

The function `MarginalProbs()` can be used to obtain some descriptive summary measures of the data:

```
MarginalProbs(Dataset = Schizo_Bin, Surr = BPRS_Bin, True = PANSS_Bin, Treat = Treat)
```

```
## $Theta_T0S0
## [1] 68.541667
##
## $Theta_T1S1
## [1] 141.61039
##
## $Freq.Cont
##
##       0   1
##   0 105  12
##   1  12  94
##
## $Freq.Exp
##
##       0   1
##   0  94   7
##   1  11 116
##
```

```
## $pi1_1_
## [1] 0.42152466
##
## $pi0_1_
## [1] 0.053811659
##
## $pi1_0_
## [1] 0.053811659
##
## $pi0_0_
## [1] 0.47085202
##
## $pi_1_1
## [1] 0.50877193
##
## $pi_1_0
## [1] 0.030701754
##
## $pi_0_1
## [1] 0.048245614
##
## $pi_0_0
## [1] 0.4122807
##
## attr(,"class")
## [1] "MarginalProbs"
```

In the output, the `Theta_T0S0` (θ_{T0S0}) and `Theta_T1S1` (θ_{T1S1}) components contain the estimated odds ratios for $S =$ clinically meaningful change on BPRS and $T =$ clinically meaningful change on PANSS in the active control and risperidone treatment groups, respectively. As it can be seen, the association between S and T is stronger in the experimental treatment group ($\widehat{\theta}_{T1S1} = 141.6104$) than in the control treatment group ($\widehat{\theta}_{T0S0} = 68.5417$).

Further, the `Freq.Cont` and `Freq.Exp` components in the output provide the frequencies for the cross-tabulation of S versus T in the control and experimental groups. For example, `Freq.Cont` shows that 12 patients had $S = 1$ and $T = 0$ in the control group. Towards the end of the output, estimates are provided for the identifiable marginal probabilities. For example, `pi1_1_` provides an estimate for $\pi_{1\cdot 1\cdot} = P(T = 1, S = 1 \mid Z = 0) = 94/223 = 0.4215$, and the other marginal probabilities are obtained in a similar way.

3.2 Sampling Γ_D: A Monte-Carlo procedure

The ICA and the SPF are functions of the parameters $\boldsymbol{\pi}$ characterizing the distribution of the vector of potential outcomes $\mathbf{Y}$. Therefore, the first step in the computation of the ICA and the SPF is the implementation of a Monte-Carlo algorithm to uniformly sample vectors $\boldsymbol{\pi}$, in the region of the parametric space that is compatible with the data, i.e., Γ_D. In addition, one may also sample a sub-region of Γ_D that has a special conceptual meaning like, for instance, the sub-region of Γ_D where monotonicity holds for both endpoints. In the *Surrogate* library, samples of $\boldsymbol{\pi}$ can be obtained using the functions `ICA.BinBin()`, `ICA.BinBin.Grid.Full()`, or `ICA.BinBin.Grid.Sample()` (for details, see the *Surrogate* manual). Due to its better numerical performance, in the present work the `ICA.BinBin.Grid.Sample()` function will be used.

The `ICA.BinBin.Grid.Sample()` function requires the user to specify the following main arguments:

- `pi1_1_=, pi0_1_=, ..., pi_0_1=` : the identifiable marginal probabilities which can be obtained using the `MarginalProbs()` function as shown earlier.
- `Monotonicity=` : the assumption that is made regarding monotonicity with:

  `Monotonicity=c("Surr.True.Endp")`: monotonicity holds for both S and T;
  `Monotonicity=c("Surr.Endp")` or `Monotonicity=c("True.Endp")`: monotonicity holds for S alone or for T alone;
  `Monotonicity=c("No")`: Monotonicity holds neither for S nor for T
  `Monotonicity=c("General")`, all four monotonicity scenarios are considered.
- `M=` : the number of runs that are conducted, i.e., the number of π vectors that are sampled. Note that when the `Monotonicity=c("General")` argument is used and thus 4 different monotonicity settings are considered, the total number of runs that are conducted is $4 * M$.

Here, a general analysis (`Monotonicity=c("General")`) is requested using $M = 10000$:

```
ICA <- ICA.BinBin.Grid.Sample(pi1_1_=0.4215, pi0_1_=0.0538, pi1_0_=0.0538,
pi_1_1=0.5088, pi_1_0=0.0307, pi_0_1=0.0482, Seed=1, Monotonicity=c("General"),
M=10000) #seed for reproducibility
```

The fitted object `ICA` contains the π vectors that are needed to compute ICA (Section 3.3) and SPF (see Section 3.5). These vectors can be obtained using the command `ICA$Pi.Vectors`. For example, the following code gives the first π vector:

```
ICA$Pi.Vectors[1:1,1:16]

##      Pi_0000      Pi_0100      Pi_0010     Pi_0001    Pi_0101      Pi_1000
## 1 0.28341818 0.0016506401 0.0040114026 0.014152335 0.17167884 0.0055954769
##      Pi_1010      Pi_1001     Pi_1110     Pi_1101     Pi_1011    Pi_1111
## 1 0.11927494 0.0078824721 0.010471947 0.027014743 0.015460923 0.27629219
##        Pi_0110    Pi_0011     Pi_0111     Pi_1100
## 1 0.0052701051 0.01070427 0.033814223 0.013307308
```

3.3 The individual causal association (ICA)

The computation of ICA using the *Surrogate* package is straightforward. Indeed, the aforementioned `ICA.BinBin.Grid.Sample()` function also computes the ICA, the odds ratios for the true endpoint T (θ_T), and the odds ratios for the surrogate endpoint S (θ_S).

The `summary()` function provides descriptive statistics for these metrics across the different monotonicity scenarios:

```
summary(ICA)

##
## Function call:
##
## ICA.BinBin.Grid.Sample(pi1_1_ = 0.4215, pi1_0_ = 0.0538, pi_1_1 = 0.5088,
##     pi_1_0 = 0.0307, pi0_1_ = 0.0538, pi_0_1 = 0.0482, Monotonicity = c("General"),
##     M = 10000, Seed = 1)
##
```

```
## # Number of valid Pi vectors
## #~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
##
## Total: 8024
##
## In the different montonicity scenarios:
##       No     True     Surr SurrTrue
##       86       55       84     7799
##
##
## # Summary of results obtained in different monotonicity scenarios
## #~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
##
## # R2_H results summary
## ~~~~~~~~~~~~~~~~~~~~
##
## Mean:
##       No     True     Surr SurrTrue
##   0.5280   0.2411   0.2695   0.1304
##
## Median:
##       No     True     Surr SurrTrue
##  0.54752  0.24385  0.26325  0.08583
##
## Mode:
##      No   True   Surr SurrTrue
##  0.5654 0.2585 0.2718  0.01309
##
## SD:
##       No     True     Surr SurrTrue
##  0.09635  0.13093  0.13832  0.13484
##
## Min:
##        No      True      Surr  SurrTrue
## 2.352e-01 3.396e-03 2.187e-02 3.086e-08
##
## Max:
##       No     True     Surr SurrTrue
##   0.6951   0.5599   0.6114   0.6322
##
##
## # Theta_T results summary
## #~~~~~~~~~~~~~~~~~~~~~~~
##
## Mean:
##       No     True     Surr SurrTrue
##    5.661      Inf   73.820      Inf
##
## Median:
##       No     True     Surr SurrTrue
##    1.205      Inf   54.582      Inf
##
## SD:
##       No     True     Surr SurrTrue
```

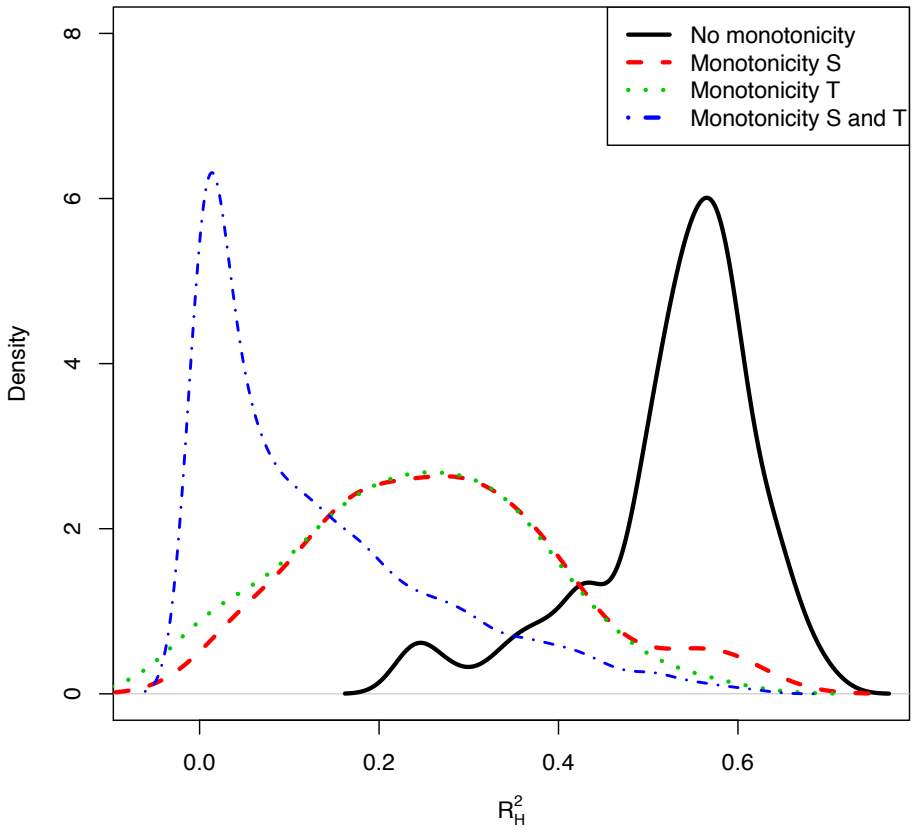
```
##    12.01      NaN    61.57      NaN
##
## Min:
##       No     True     Surr SurrTrue
##   0.0278      Inf  18.5317      Inf
##
## Max:
##       No     True     Surr SurrTrue
##     86.5      Inf    378.8      Inf
##
##
## # Theta_S results summary
## #~~~~~~~~~~~~~~~~~~~~~~
##
## Mean:
##       No     True     Surr SurrTrue
##    6.523   64.972      Inf      Inf
##
## Median:
##       No     True     Surr SurrTrue
##    1.219   56.829      Inf      Inf
##
## SD:
##       No     True     Surr SurrTrue
##    16.12    44.08      NaN      NaN
##
## Min:
##       No     True     Surr SurrTrue
##  0.01921 17.23491      Inf      Inf
##
## Max:
##       No     True     Surr SurrTrue
##    117.6    232.8      Inf      Inf
```

The first part of the output shows the number of valid vectors $\boldsymbol{\pi}$ (vectors in Γ_D) obtained in the analysis. As it can be seen, the 40000 runs of the algorithm led to 8024 vectors $\boldsymbol{\pi}$ compatible with the data. These valid vectors are subsequently used to compute R^2_H, θ_S, and θ_T. For these metrics, the means, medians, mode, SD, minimum and maximum values, obtained under the different monotonicity scenarios, are provided. The `No`, `True`, `Surr`, and `SurrTrue` labels depict the results that are obtained in the no monotonicity, monotonicity for T alone, monotonicity for S alone, and monotonicity for both S and T scenarios, respectively.

The largest estimates for the measures of central tendency of ICA were obtained when no monotonicity was assumed, with $\widehat{R}^2_H$ mean = 0.5280, median = 0.5475, mode = 0.5654 (SD = 0.0964, range $[0.2352; 0.6951]$). Furthermore, the lowest estimates were obtained when monotonicity was assumed for both S and T, with $\widehat{R}^2_H$ mean = 0.1304, median = 0.0858, mode = 0.0131 (SD = 0.1348, range $[0.0001; 0.6322]$). Finally, when monotonicity was assumed for S alone and for T alone the estimates of the measures of central tendency lied between the previous ones.

The density functions for R^2_H across different monotonicity scenarios can be obtained using the following command:

```
plot(ICA, ylim=c(0, 8))
```



As it can be seen, when monotonicity is assumed for both S and T (blue line in the figure) small values for R^2_H are much more supported than large values, whereas when monotonicity is not assumed (black line in the figure) large values received more support. When monotonicity is assumed for only one endpoint the frequency densities lie between the ones obtained in the previous two cases and, here again, smaller values are more supported than large ones.

Assessing surrogacy under such a high level of uncertainty is obviously challenging. As the previous analyses clearly show, the results are rather sensitive to the unverifiable monotonicity assumptions. Nonetheless, domain specific knowledge can sometimes shed light on the plausibility of these competing assumptions. In the next section this idea is further illustrated.

3.4 Exploring the plausibility of different scenarios

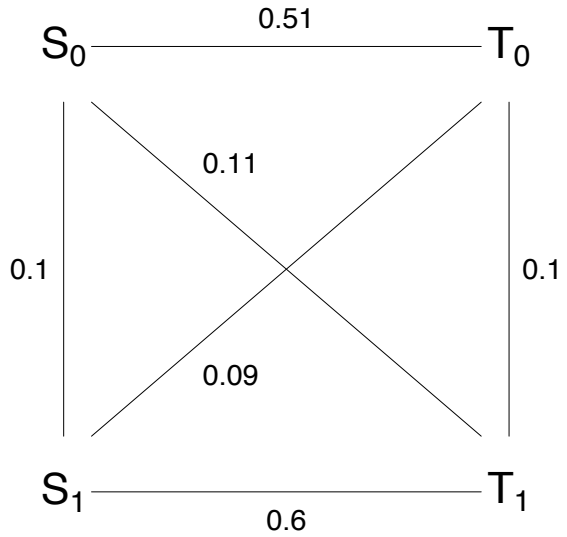
Causal diagrams All the previous frequency densities emanate from vectors π that are equally compatible with the data at hand and, hence, the frequency densities are themselves equally compatible with the data. Therefore, based solely on the data one cannot discriminate between the scenario in which large values of R^2_H received more support, basically the setting in which monotonicity holds for neither endpoint, and the other scenarios in which smaller values are more supported, i.e, the settings in which monotonicity holds for at least one endpoint.

However, in some situations, domain specific knowledge can be used to evaluate the plausibility of the different scenarios. The function `CausalDiagramBinBin()` may play an important role in this context. This function shows a causal diagram that depicts the median of the informational coefficients of association (r_h^2) or odds ratios, describing the association structure for the counterfactual vector $\mathbf{Y} = (T_0, T_1, S_0, S_1)'$. The function can also be used to describe the association structure of $\mathbf{Y}$ in a specified subgroup defined by the values of the ICA. The following arguments are required:

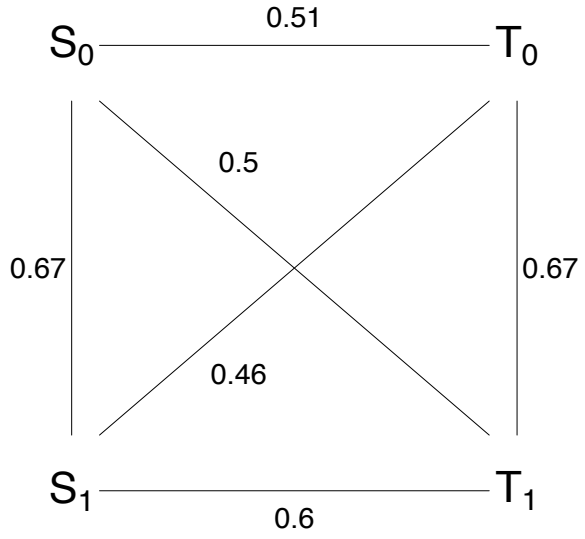
- `x=` : a fitted object of class `ICA.BinBin`.
- `Values=` : specifies whether the median informational coefficients of correlation (`Values="Corrs"`) or median odds ratios (`Values="ORs"`) between the counterfactuals should be depicted.
- `Min=`, `Max=` : the minimum and maximum values for $\widehat{R}_H^2$ that should be considered.
- `Monotonicity=` : the monotonicity scenario that should be considered.
- `Histograms.Correlations` : specifies whether histograms of the informational coefficients of association R_H^2 between the counterfactuals should be provided. Defaults to `Histograms.Correlations=FALSE`.

For example, the following commands provide causal diagrams that depict the median informational coefficients of association between the counterfactuals under the four different monotonicity scenarios:

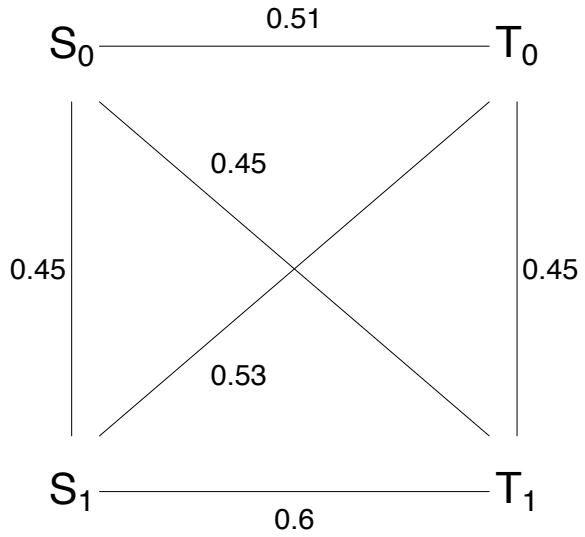
```
CausalDiagramBinBin(x=ICA, Monotonicity="No")

## Note. The figure is based on 86 observations.
```



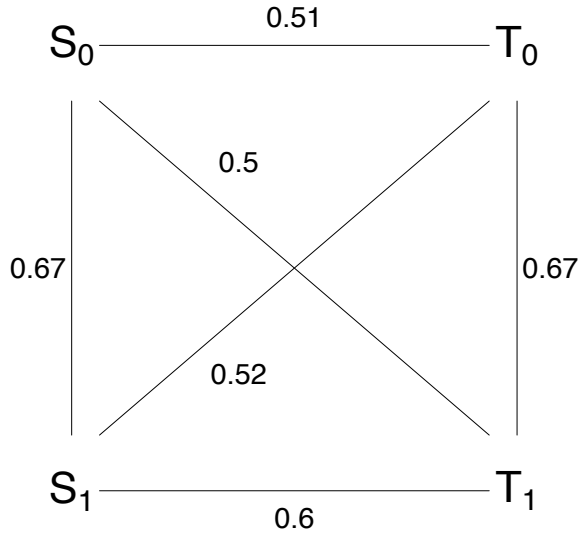

```
CausalDiagramBinBin(x=ICA, Monotonicity="Surr.Endp")

## Note. The figure is based on 84 observations.
```




```
CausalDiagramBinBin(x=ICA, Monotonicity="True.Endp")

## Note. The figure is based on 55 observations.
```




```
CausalDiagramBinBin(x=ICA, Monotonicity="Surr.True.Endp")

## Note. The figure is based on 7799 observations.
```



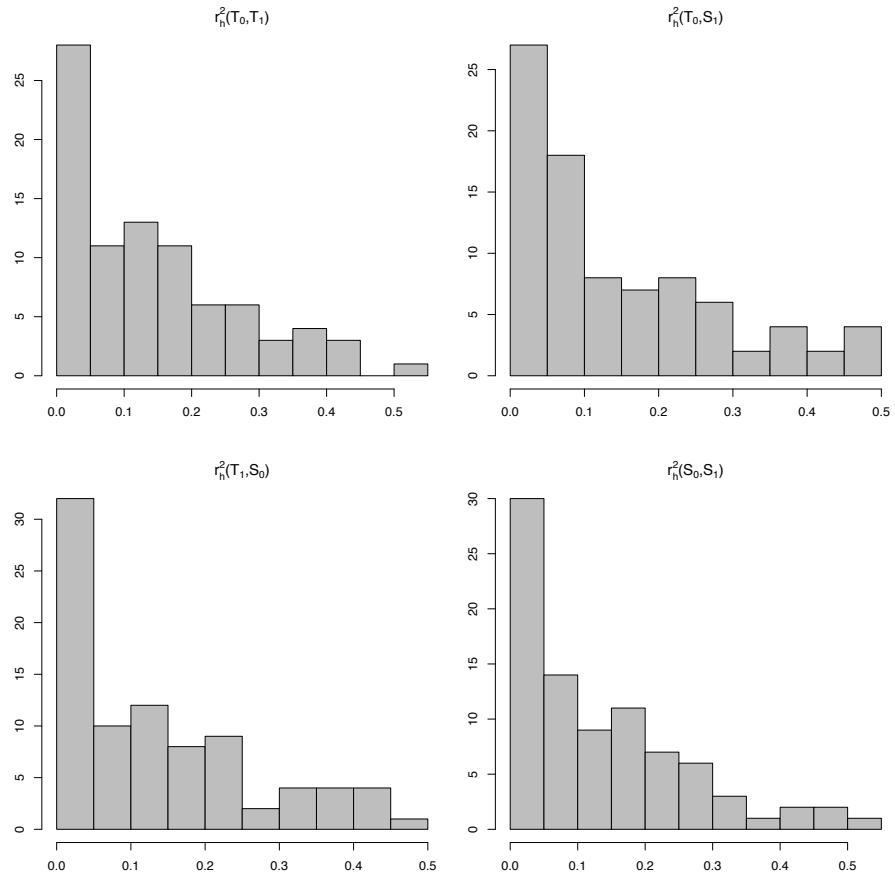
In these diagrams, the two horizontal lines depict the identifiable informational coefficients of association between S and T in the two treatment conditions, i.e., $\widehat{r}_h^2(S_0,\, T_0) = 0.51$ and $\widehat{r}_h^2(S_1,\, T_1) = 0.60$. Essentially, these coefficients quantify the association between the surrogate and the true endpoint in both treatment groups and can be interpreted along the lines presented in Alonso *et al.* (2016a).

The other four non-horizontal lines depict the medians of the unidentified informational coefficients of association between the counterfactuals. When monotonicity is not assumed (first causal diagram), the median informational association between the potential outcomes for the true and surrogate endpoints are small, i.e., $\widehat{r}_h^2(S_0,\, S_1) = \widehat{r}_h^2(T_0,\, T_1) = 0.10$. This means that a patient's outcome on BPRS/PANSS in the active control condition (S_0/T_0)conveys little information on the patient's outcome on BPRS/PANSS in the experimental treatment condition (S_1/T_1). Given that the treatments under study are similar and S_0, S_1 (and also T_0, T_1) are repeated measurements in the same patient, this weak association may be considered counter-intuitive. Further, the other median informational associations $\widehat{r}_h^2(S_0,\, T_1) = 0.11$ and $\widehat{r}_h^2(S_1,\, T_0) = 0.09$ are also low. Since the BPRS is a sub-scale of the more complex PANSS scale, one would also expect a certain level of association between these potential outcomes and independence is again counter-intuitive.

When monotonicity is assumed for S alone, T alone, or for both S and T (the other three causal diagrams), the median informational associations between the potential outcomes are substantially larger. For example, when monotonicity is assumed for S, the median $\widehat{r}_h^2(S_0,\, S_1) = \widehat{r}_h^2(T_0,\, T_1) = 0.67, \widehat{r}_h^2(S_0,\, T_1) = 0.50$ and $\widehat{r}_h^2(S_1,\, T_0) = 0.46$. As was discussed in the previous paragraph, this pattern of association between the potential outcomes seems to be more compatible with our biological expectations.

The non-horizontal lines in the causal diagrams shown above depict the *median* of the informational coefficients of association. It can also be informative to explore the *whole distribution* of the informational coefficients of association. These distributions can be obtained by adding the `Histograms.Correlations = TRUE` argument in the `CausalDiagramBinBin()` function call. For example, the histograms of the informational coefficients of association that are obtained in the no monotonicity scenario can be obtained using the following command:

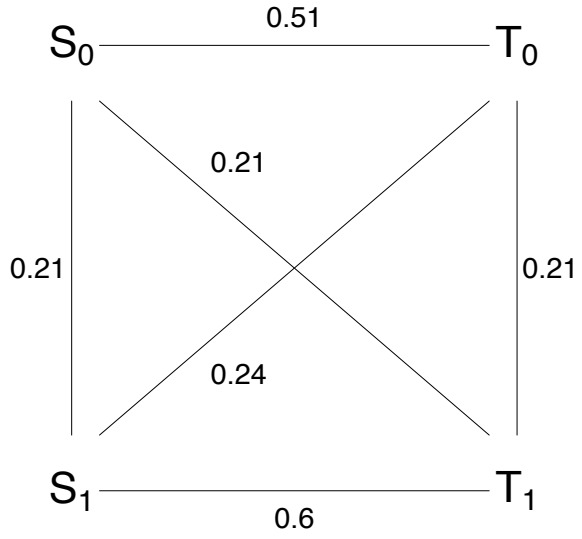
```
CausalDiagramBinBin(x=ICA, Monotonicity="No", Histograms.Correlations=TRUE)
```



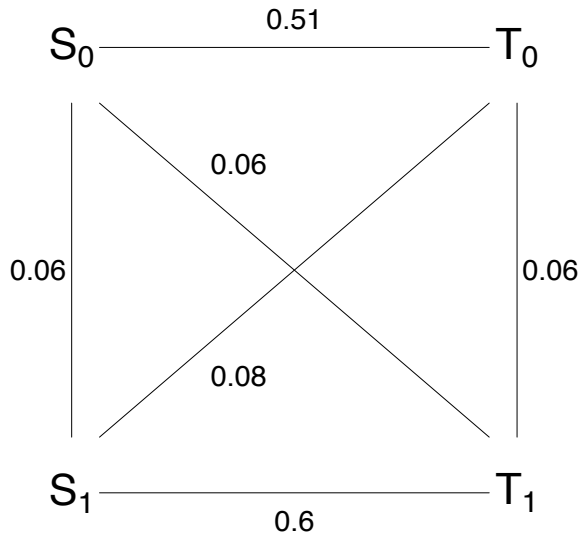
These figures show that most of the informational coefficients of association that are compatible with the observed data are close to zero. It is also interesting to note that even though the medians of $\hat{r}^2_h(S_0,\, S_1)$ and $\hat{r}^2_h(T_0,\, T_1)$ are always the same (as shown in the causal diagrams), their actual distributions differ (though they are similar).

Further, a closer look to the frequency density obtained under the no monotonicity assumption can be insightful as well. Indeed, this is the only scenario in which large values of R^2_H are more supported than small values. The following command can be used to obtain causal diagrams under the no monotonicity scenario that are compatible with $R^2_H \leq 0.50$ and $R^2_H \geq 0.50$:

```
CausalDiagramBinBin(x=ICA, Monotonicity="No", Min = 0, Max = .5)

## Note. The figure is based on 23 observations.
```




```
CausalDiagramBinBin(x=ICA, Monotonicity="No", Min = 0.5, Max = 1)

## Note. The figure is based on 63 observations.
```



As can be seen, the larger R^2_H values seem to happen primarily when all the unidentifiable associations are rather small (all $\widehat{r}^2_h \leq 0.08$). If one renders these low associations biologically implausible, then also in this scenario small values of R^2_H should be taken as more biologically meaningful.

Clearly, there will always be a certain level of subjectivity in this type of qualitative analysis and expert opinion may be of great value when interpreting the previous diagrams in order to evaluate the biological plausibility of the different monotonicity assumptions.

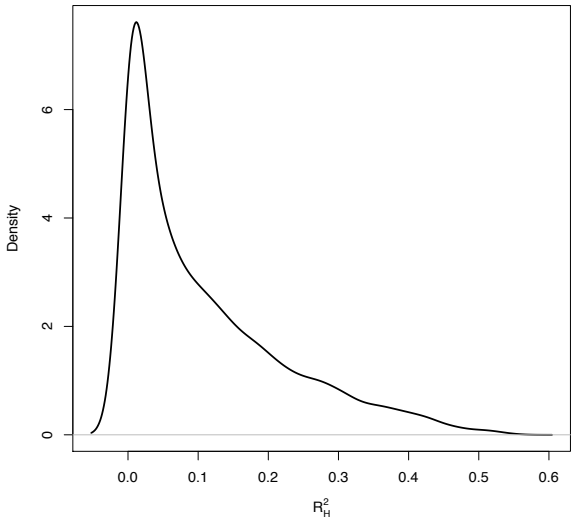
Setting biologically plausible restrictions on the counterfactual correlations The function `ICA.BinBin.CounterAssum` (ICA in the Binary-Binary setting where the Counterfactuals are Assumed

to fall within a certain range) is also useful to explore sub-regions of Γ_D that are of special interest. The function requires the user to specify the following main arguments:

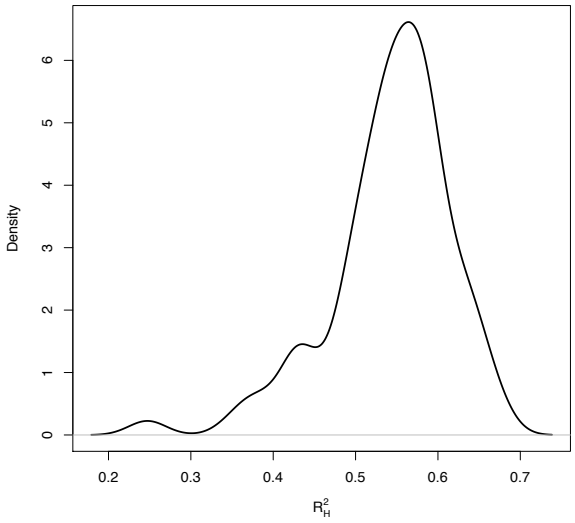
- `x=` : a fitted object of class `ICA.BinBin`.
- `r2_h_S0S1_min=`, `r2_h_S0S1_max=` : the minimum and maximum values to be considered for $r^2_h(S_0, S_1)$, and similarly for the other informational coefficients of association.
- `Monotonicity=` : the monotonicity scenario that should be considered.
- `Type=` : the type of plot that should be provided (i.e., `Type="Freq"`, `Type="Density"`, or `Type="All.Densities"`).

For example, let us assume that, based on expert opinion, the sub-region of Γ_D where $\rho_{S0S1} > 0.5$, $\rho_{S0T1} > 0.4$, $\rho_{T0T1} > 0.5$, and $\rho_{T0S1} > 0.4$ is more biologically plausible. Descriptives of the R^2_H values and density plots that are obtained in the biologically more plausible and biologically less plausible scenarios can be obtained using the following commands:

```
# Biologically plausible scenario
ICA.BinBin.CounterAssum(x = ICA, r2_h_S0S1_min = .5, r2_h_S0S1_max = 1,
r2_h_S0T1_min = .4, r2_h_S0T1_max = 1, r2_h_T0T1_min = .5, r2_h_T0T1_max = 1,
r2_h_T0S1_min = .4, r2_h_T0S1_max = 1, Monotonicity = "General",
Type = "Density")


##
##
## Summary measures for R2_H (in the subgroup of results where the counterfactual
## correlations fall within prespecified ranges
## ###############################################################################
##
##
## # R2_H results summary
## #~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
##
## Mean (SD) R2_H: 0.1080 (0.1140)  [min: 0.0000; max: 0.5521]
## Mode R2_H:  0.0119
##
## Quantiles of the R2_H distribution:
##
##        5%       10%       20%       50%       80%       90%       95%
## 0.0005571 0.0022302 0.0094113 0.0685515 0.1972203 0.2817226 0.3514906
##
##
## Note. The figure is based on 7036 observations.
```




```r
# Biologically implausible scenario
ICA.BinBin.CounterAssum(x = ICA, r2_h_S0S1_min = 0, r2_h_S0S1_max = 0.5,
r2_h_S0T1_min = 0, r2_h_S0T1_max = 0.4, r2_h_T0T1_min = 0, r2_h_T0T1_max = 0.5,
r2_h_T0S1_min = 0, r2_h_T0S1_max = 0.4, Monotonicity = "General",
Type = "Density")

##
##
## Summary measures for R2_H (in the subgroup of results where the counterfactual
## correlations fall within prespecified ranges
## ################################################################################
##
##
## # R2_H results summary
## #~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
##
## Mean (SD) R2_H: 0.5421 (0.0742)  [min: 0.2472; max: 0.6703]
## Mode R2_H:  0.5644
##
## Quantiles of the R2_H distribution:
##
##     5%    10%    20%    50%    80%    90%    95%
## 0.4176 0.4352 0.4993 0.5517 0.5909 0.6291 0.6390
##
##
## Note. The figure is based on 78 observations.
```



The output shows that the $\widehat{R}^2_H$ is substantially higher in the biologically less plausible scenario ($\widehat{R}^2_H$ mean = 0.5421, median = 0.5517, mode = 0.5644, and 95% of the $\widehat{R}^2_H > 0.4176$) than in the biologically more plausible scenario ($\widehat{R}^2_H$ mean = 0.1080, median = 0.0686, mode = 0.0119, and 95% of the $\widehat{R}^2_H < 0.3515$).

3.5 The surrogate predictive function (SPF)

It was observed, in Section 3.3, that monotonicity had a substantial impact on the results, i.e., ICA tended to be substantially higher in the no monotonicity scenario compared to the scenarios where monotonicity was assumed for S alone, for T alone, and for both S and T. The ICA, as given by R^2_H, can be interpreted as a measure of prediction accuracy, with values close to zero indicating independent individual causal treatment effects on S and T (no meaningful prediction can be made) and values close to one giving evidence of a deterministic relationship between ΔT and ΔS (prediction without error).

In the following sections SPF will be estimated in the scenarios where monotonicity holds for neither endpoint (only setting in which larger values of ICA are more supported) and when monotonicity is valid for S alone, T alone and for both S and T (low ICA values more supported). The use of SPF in conjunction with R^2_H can help to assess the surrogate effect predictive value. In fact, while R^2_H offers a general quantification of the surrogate predictive capacity, SPF zooms in to offer a more detailed view on how ΔT and ΔS are related.

3.5.1 Analysis in the no monotonicity scenario

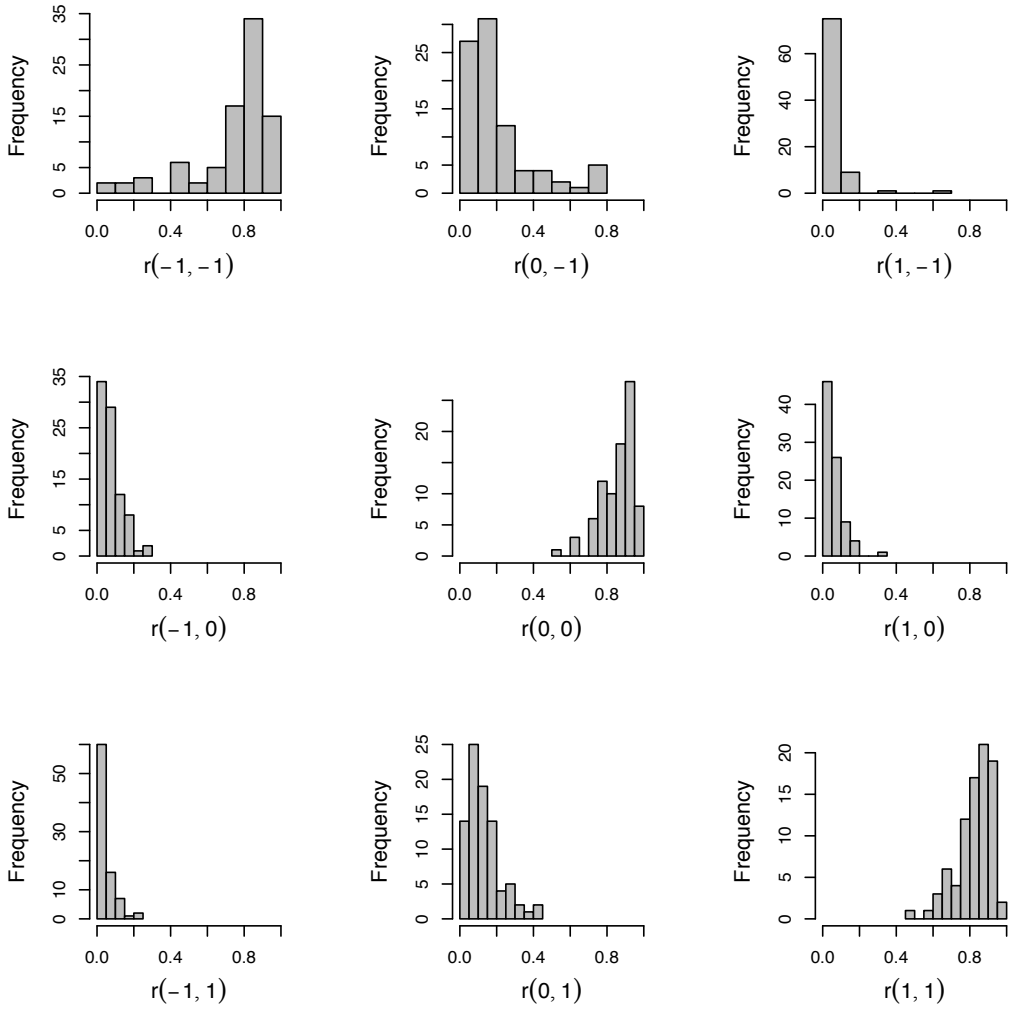
The function `SPF.BinBin` (Surrogate Predictive Function for Binary S and Binary T) computes SPF using the sensitivity analysis strategy proposed by Alonso *et al.* (2016b). The function requires the user to specify a fitted object of class `ICA.BinBin` which contains the $\boldsymbol{\pi}$ vectors that are needed to determine SPF (see Section 3.2). To obtain the SPF under the assumption of no monotonicity, the following commands can be used:

```
ICA_No <- ICA.BinBin.Grid.Sample(pi1_1_=0.4215, pi0_1_=0.0538, pi1_0_=0.0538,
pi_1_1=0.5088, pi_1_0=0.0307, pi_0_1=0.0482, Seed=1,
Monotonicity=c("No"), M=10000) #seed for reproducibility

SPF_No <- SPF.BinBin(ICA_No)
```

```
summary(SPF_No)

##
## Function call:
##
## SPF.BinBin(x = ICA_No)
##
##
## Total number of valid Pi vectors
## ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## 86
##
##
## SPF Descriptives
## ~~~~~~~~~~~~~~~~
## r_min1_min1     Mean: 0.7484;  Median: 0.82509;  Mode: 0.87114;  SD: 0.21717
##                        Min: 0.058457; Max: 0.97168; 95% CI = [0.13888; 0.95012]
##
## r_0_min1        Mean: 0.2033;  Median: 0.13568;  Mode: 0.1034;  SD: 0.18458
##                        Min: 0.00467; Max: 0.7679; 95% CI = [0.028757; 0.72956]
##
## r_1_min1        Mean: 0.048296;  Median: 0.023475;  Mode: 0.014946;  SD: 0.083193
##                        Min: 0.0001159; Max: 0.60665; 95% CI = [0.00056563; 0.19158]
##
## r_min1_0        Mean: 0.078138;  Median: 0.05819;  Mode: 0.042112;  SD: 0.056138
##                        Min: 0.0059661; Max: 0.25208; 95% CI = [0.0092982; 0.20547]
##
## r_0_0           Mean: 0.85965;  Median: 0.89062;  Mode: 0.91531;  SD: 0.087427
##                        Min: 0.54755; Max: 0.97051; 95% CI = [0.63161; 0.95874]
##
## r_1_0           Mean: 0.062216;  Median: 0.047302;  Mode: 0.039454;  SD: 0.048094
##                        Min: 0.0037408; Max: 0.31073; 95% CI = [0.0078327; 0.16727]
##
## r_min1_1        Mean: 0.044326;  Median: 0.030975;  Mode: 0.022397;  SD: 0.045417
##                        Min: 0.00068337; Max: 0.23243; 95% CI = [0.0012776; 0.17517]
##
## r_0_1           Mean: 0.13028;  Median: 0.11176;  Mode: 0.079046;  SD: 0.089613
##                        Min: 0.0044017; Max: 0.42858; 95% CI = [0.020308; 0.35413]
##
## r_1_1           Mean: 0.8254;  Median: 0.84728;  Mode: 0.88384;  SD: 0.099755
##                        Min: 0.45564; Max: 0.97397; 95% CI = [0.6032; 0.94632]
```

A plot of the SPF histograms can be obtained by applying the `plot()` function to the fitted object `SPF_No`:

```
plot(SPF_No)
```



The output of the `summary()` function shows that the 10000 runs of the algorithm led to 86 valid $\boldsymbol{\pi}$ vectors, i.e., vector in the sub-region of Γ_D where monotonicity holds neither for T nor for S. Further, descriptives like the mean, median, and mode for each of the $r(i,j) = P(\Delta T = i|\Delta S = j)$ are provided. In addition, the output of the `plot()` function shows the histograms for SPF ($r(i,j)$). A discussion of the output is provided in Alonso *et al.* (2016b).

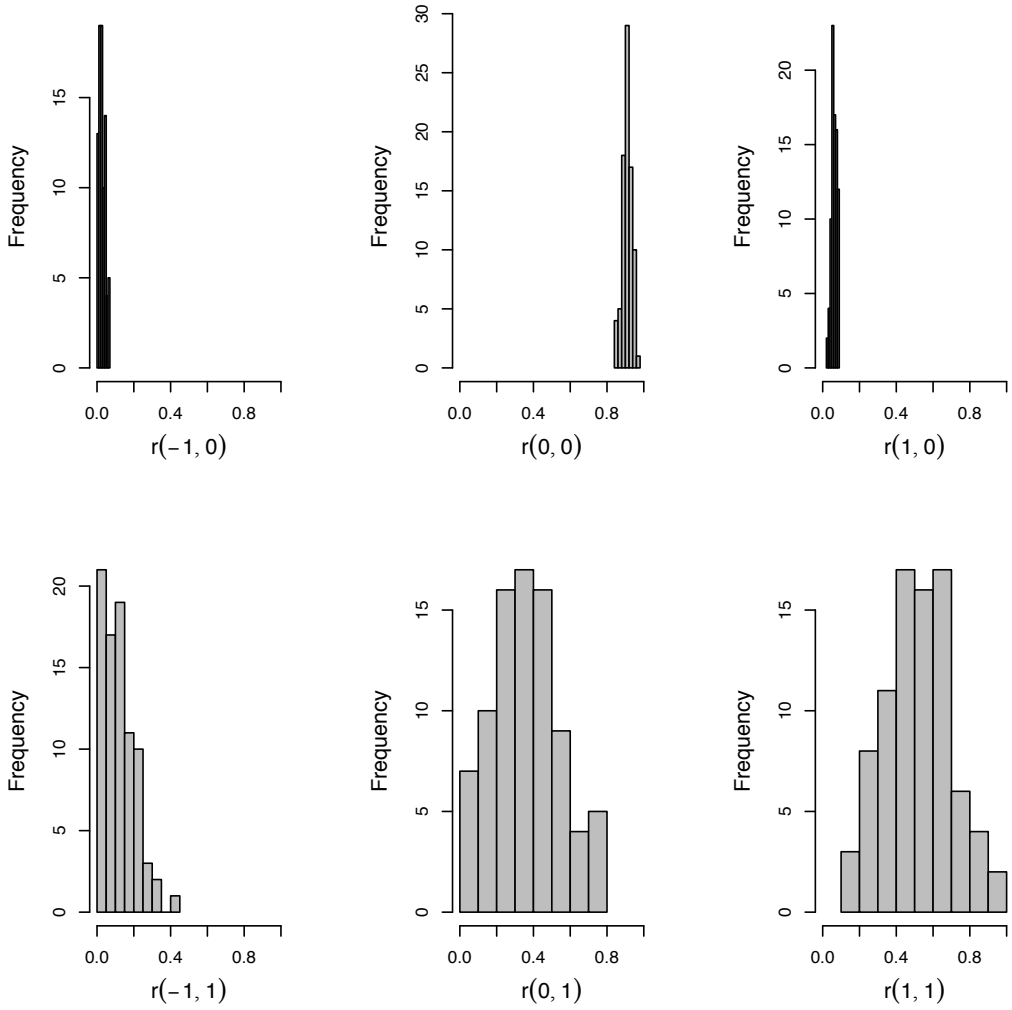
3.5.2 Analysis in the monotonicity for S scenario

To obtain the SPF under the assumption of monotonicity for S, the following commands can be used:

```
ICA_S <- ICA.BinBin.Grid.Sample(pi1_1_=0.4215, pi0_1_=0.0538, pi1_0_=0.0538,
pi_1_1=0.5088, pi_1_0=0.0307, pi_0_1=0.0482, Seed=1,
Monotonicity=c("Surr.Endp"), M=10000) #seed for reproducibility

SPF_S <- SPF.BinBin(ICA_S)

summary(SPF_S)

##
```

```
## Function call:
##
## SPF.BinBin(x = ICA_S)
##
##
## Total number of valid Pi vectors
## ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## 84
##
##
## SPF Descriptives
## ~~~~~~~~~~~~~~~~
## r_min1_0        Mean: 0.028009;  Median: 0.024862;  Mode: 0.017337;  SD: 0.016733
##                        Min: 0.0028209; Max: 0.06698; 95% CI = [0.005466; 0.065813]
##
## r_0_0           Mean: 0.90912;  Median: 0.90736;  Mode: 0.9045;  SD: 0.024644
##                        Min: 0.84563; Max: 0.9715; 95% CI = [0.85094; 0.94962]
##
## r_1_0           Mean: 0.062867;  Median: 0.061211;  Mode: 0.05893;  SD: 0.015146
##                        Min: 0.020086; Max: 0.089875; 95% CI = [0.03232; 0.087476]
##
## r_min1_1        Mean: 0.12134;  Median: 0.11724;  Mode: 0.051222;  SD: 0.090701
##                        Min: 0.00045333; Max: 0.44463; 95% CI = [0.0031818; 0.32532]
##
## r_0_1           Mean: 0.36332;  Median: 0.34078;  Mode: 0.32548;  SD: 0.17894
##                        Min: 0.051993; Max: 0.76668; 95% CI = [0.073257; 0.72383]
##
## r_1_1           Mean: 0.51534;  Median: 0.50723;  Mode: 0.47828;  SD: 0.18378
##                        Min: 0.1018; Max: 0.92469; 95% CI = [0.18044; 0.88129]
```

```
plot(SPF_S)
```



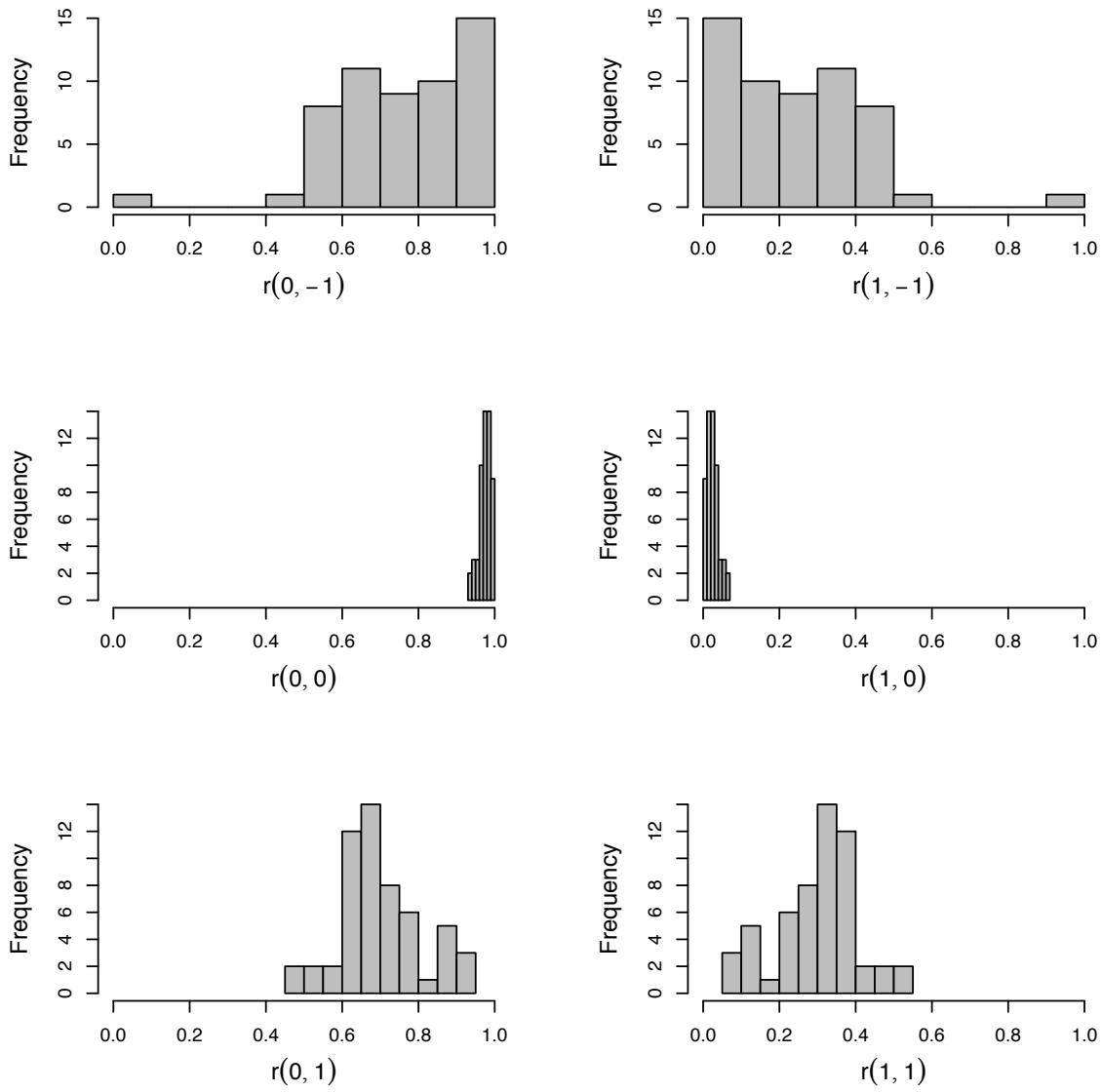
Notice that the previous outputs do not show estimates for $r(i, j = -1)$, as the probabilities of these events are 0 when monotonicity for S is assumed. A discussion of the output is provided in Alonso *et al.* (2016b).

3.5.3 SPF assuming monotonicity for T

To obtain the SPF under the assumption of monotonicity for T, the following commands can be used:

```
ICA_T <- ICA.BinBin.Grid.Sample(pi1_1_=0.4215, pi0_1_=0.0538, pi1_0_=0.0538,
pi_1_1=0.5088, pi_1_0=0.0307, pi_0_1=0.0482, Seed=1,
Monotonicity=c("True.Endp"), M=10000) #seed for reproducibility

SPF_T <- SPF.BinBin(ICA_T)

summary(SPF_T)

##
## Function call:
##
```

```
## SPF.BinBin(x = ICA_T)
##
##
## Total number of valid Pi vectors
## ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## 55
##
##
## SPF Descriptives
## ~~~~~~~~~~~~~~~~
## r_0_min1        Mean: 0.75868;  Median: 0.77398;  Mode: 0.89699;  SD: 0.17762
##                        Min: 0.087598; Max: 0.9997; 95% CI = [0.46344; 0.99292]
##
## r_1_min1        Mean: 0.24132;  Median: 0.22602;  Mode: 0.10301;  SD: 0.17762
##                        Min: 0.00030333; Max: 0.9124; 95% CI = [0.0070815; 0.53656]
##
## r_0_0           Mean: 0.97411;  Median: 0.97673;  Mode: 0.97997;  SD: 0.015363
##                        Min: 0.93769; Max: 0.99771; 95% CI = [0.93961; 0.99633]
##
## r_1_0           Mean: 0.025885;  Median: 0.023271;  Mode: 0.020026;  SD: 0.015363
##                        Min: 0.0022875; Max: 0.062315; 95% CI = [0.0036691; 0.060389]
##
## r_0_1           Mean: 0.69919;  Median: 0.6848;  Mode: 0.66413;  SD: 0.10829
##                        Min: 0.45411; Max: 0.93901; 95% CI = [0.50732; 0.90863]
##
## r_1_1           Mean: 0.30081;  Median: 0.3152;  Mode: 0.33587;  SD: 0.10829
##                        Min: 0.060992; Max: 0.54589; 95% CI = [0.091367; 0.49268]

plot(SPF_T)
```



In line with the results detailed in Section 3.5.2, it can be concluded that a lack of treatment effect on S will strongly suggest a lack of treatment effect on T, however, a positive or negative impact of the treatment on S coveys limited information on the potential effect of the treatment on T.

3.5.4 SPF assuming monotonicity for both S and T

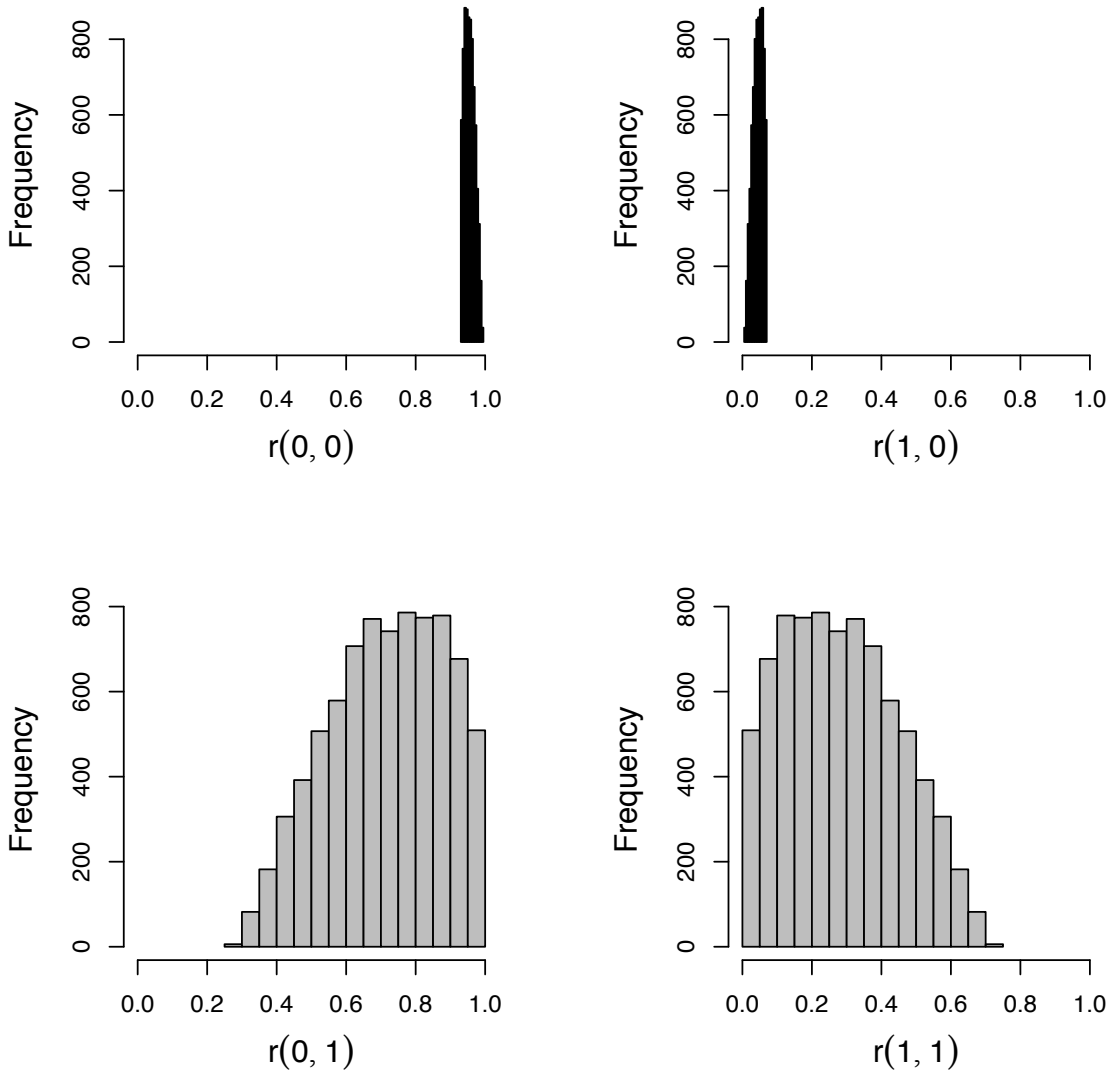
The following commands can be used to obtain the SPF under the assumption of monotonicity for both S and T:

```
ICA_ST <- ICA.BinBin.Grid.Sample(pi1_1_=0.4215, pi0_1_=0.0538, pi1_0_=0.0538,
pi_1_1=0.5088, pi_1_0=0.0307, pi_0_1=0.0482, Seed=1,
Monotonicity=c("Surr.True.Endp"), M=10000) #seed for reproducibility

SPF_ST <- SPF.BinBin(ICA_ST)
```

```r
summary(SPF_ST)
```

```
##
## Function call:
##
## SPF.BinBin(x = ICA_ST)
##
##
## Total number of valid Pi vectors
## ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## 7799
##
##
## SPF Descriptives
## ~~~~~~~~~~~~~~~~
## r_0_0           Mean: 0.95557;  Median: 0.95451;  Mode: 0.94864;  SD: 0.014831
##                        Min: 0.93009; Max: 0.99301; 95% CI = [0.93191; 0.9851]
##
## r_1_0           Mean: 0.044435;  Median: 0.045489;  Mode: 0.051358;  SD: 0.014831
##                        Min: 0.0069928; Max: 0.069907; 95% CI = [0.014903; 0.068091]
##
## r_0_1           Mean: 0.71364;  Median: 0.72548;  Mode: 0.79145;  SD: 0.1667
##                        Min: 0.2928; Max: 0.99994; 95% CI = [0.38171; 0.97954]
##
## r_1_1           Mean: 0.28636;  Median: 0.27452;  Mode: 0.20855;  SD: 0.1667
##                        Min: 5.7e-05; Max: 0.7072; 95% CI = [0.020461; 0.61829]
```

```r
plot(SPF_ST)
```

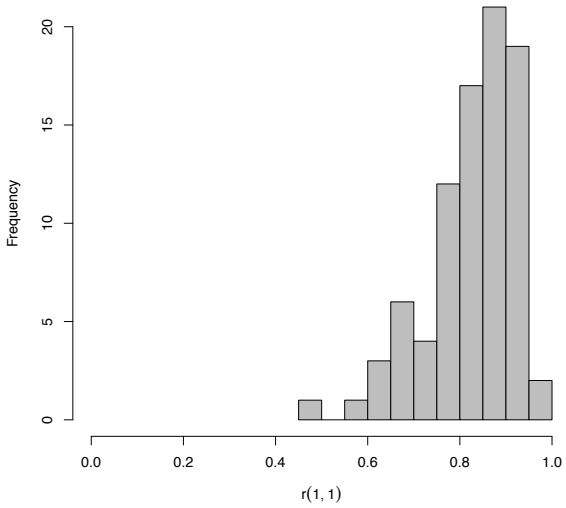


In line with the results detailed in Section 3.5.2, the results indicate that a lack of treatment effect on S is strongly indicative of a lack of treatment effect on T, but a positive impact of the treatment on S cannot be interpreted as strong evidence that there will also be a positive impact on T.

3.6 Additional graphical tools to explore the surrogate predictive function

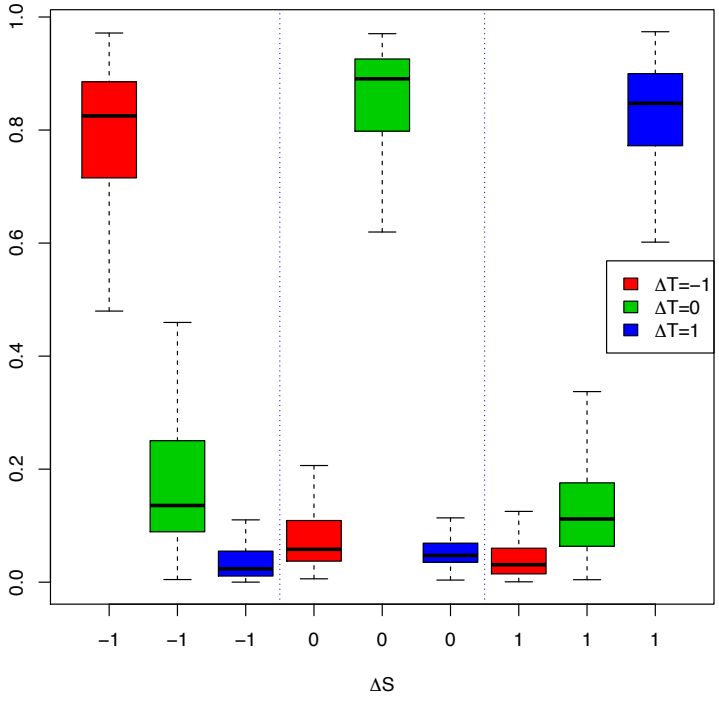
When the `plot()` function is applied to a fitted object of class `SPF.BinBin`, histograms of $r(i,j)$ are provided by default (see previous figures). Other types of plots can also be requested (for details, see the *Surrogate* package manual). For example, a histogram for a particular $r(i,j)$ of interest, e.g., $r(1,1)$, can be requested by specifying the `Type="Histogram"` and `Specific.Pi="r_1_1"` arguments in the `plot()` call. Here, we request such a figure in the no monotonicity scenario:

```
plot(SPF_No, Type="Histogram", Specific.Pi="r_1_1")
```



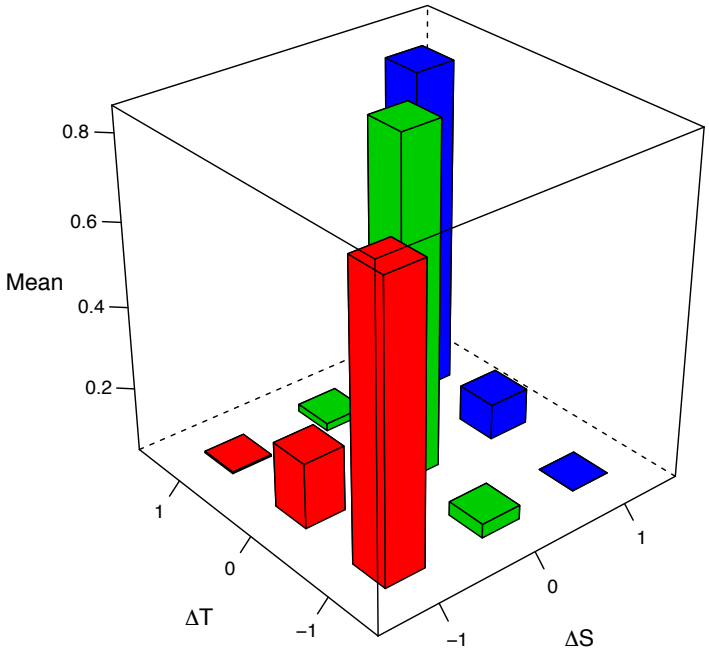
In addition, it is also possible to request box-plots for $r(i,j)$ by using the `Type="Box.Plot"` argument in the `plot()` call:

```
plot(SPF_No, Type="Box.Plot", Legend.Pos="right")
```



Further, line plots and 3D plots that depict the means, medians or modes of the $r(i,j)$ can be obtained by using the `Type="Lines.Mean"`, `Type="Lines.Median"`, `Type="Lines.Mode"`, `Type="3D.Mean"`, `Type="3D.Median"` or `Type="3D.Mode"` arguments. For example, a 3D plot of the means can be obtained using the following command:

```
plot(SPF_No, Type="3D.Mean")
```



Finally, 3D spinning plots that can be freely rotated on the screen can be obtained by using the `plot(SPF_No, Type="3D.Spinning.Mean")`, `plot(SPF_No, Type="3D.Spinning.Median")`, or `plot(SPF_No, Type="3D.Spinning.Mode")` commands.

4 Impact of ignoring the sampling variability in $\hat{b}$: Simulation study

4.1 Simulation design

The methodology proposed in Alonso *et al.* (2016b) characterizes Γ_D using the estimated components of $\boldsymbol{b}$ and, consequently, the sampling variability of these estimates is 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. In the present section, a simulation study is carried out to evaluate this issue, i.e., the impact of using $\widehat{\mathbf{b}}$ instead of $\mathbf{b}$, on the assessment of the ICA and SPF.

Table 1 shows the two scenarios considered for the identifiable marginal probabilities contained in $\mathbf{b}$. Notice that, in both scenarios, the surrogate and true endpoint are associated in the control and treated groups. Actually, in practice, the presence of an association between the putative surrogate and the true endpoint is often taken as a prerequisite for surrogacy and, therefore, we did not consider settings in which both endpoints were independent. In scenario 1 both endpoints are moderately associated with

Table 1: Different scenarios (true marginal probabilities) that were used to simulate the data.

Scenario 1								Scenario 2							
$Z=0$				$Z=1$				$Z=0$				$Z=1$			
		T				T				T				T	
		0	1			0	1			0	1			0	1
S	0	0.30	0.20	S	0	0.30	0.20	S	0	0.45	0.05	S	0	0.45	0.05
	1	0.20	0.30		1	0.20	0.30		1	0.05	0.45		1	0.05	0.45

$\theta_{ST|Z} = 2.25$. Scenario 2 represents the more extreme and probably more unrealistic setting in which both endpoints are almost deterministically related in both treatment groups, with $P(T = S|Z) = 0.9$ and $\theta_{ST|Z} = 81$. Notice that, even though it may be unlikely in practice, scenario 2 is still methodologically and conceptually interesting. Notice also that, given that BPRS is a subscale of PANSS, in the case study the association between the surrogate and true endpoint is similar to the one considered in scenario 2. Furthermore, five sample sizes were evaluated, namely, $N = 50$, 100, 300, 600 and 1000 patients. For each sample-size-scenario combination, 250 data sets were generated using draws from a multinomial distribution. Thus, in total, $2,500$ data sets were obtained and in each of these data sets $\widehat{\mathbf{b}}$ was determined.

Finally, the ICA and SPF were assessed using $\mathbf{b}$ ($\text{ICA}_{\mathbf{b}}$, $\text{SPF}_{\mathbf{b}}$), i.e., the true values given in Table 1, and the estimated values $\widehat{\mathbf{b}}$ ($\text{ICA}_{\widehat{\mathbf{b}}}$, $\text{SPF}_{\widehat{\mathbf{b}}}$) as the input for the proposed algorithm. The Monte Carlo procedure was implemented using $M = 50,000$ runs and assuming no monotonicity.

Main outcomes of interest: The main goal of the simulation study was the assessment of the bias induced by replacing $\mathbf{b}$ by $\widehat{\mathbf{b}}$ when analyzing the data. Therefore, the relative ICA bias, computed as $E\left[\left(\text{ICA}_{\widehat{\mathbf{b}}} - \text{ICA}_{\mathbf{b}}\right) / \text{ICA}_{\mathbf{b}}\right]$ was one of the studied outcomes. A similar outcome was also considered for the SPF.

4.2 Simulation results

Tables 2 and 3 display the results obtained in scenarios 1 and 2, respectively. With respect to the ICA, the results showed that the biased induced by ignoring the sample variability is mostly negligible. Only when the sample size was rather small, i.e., $N = 50$ patients, certain degree of bias was observed, but it never exceeded 15%. Importantly, for a sample size smaller than the one of the case study, i.e., $N = 300$, the relative bias was only about 1.3% in both scenarios.

With respect to the SPF, the relative bias in scenario 1 was always less than about 4% for samples of size $N = 100$ and always less than 2% for samples of size $N = 300$ or larger. Interestingly, in scenario 2, although the relative bias was generally small, for $r\,(-1,\, 1)$ and $r\,(1,\, -1)$ large relatively biases were observed. For example, when $N = 300$, the relative bias for these values was about 11%. As expected, for sample sizes larger than $N = 600$, the relative bias was much smaller. Actually, for $N = 1000$ the relative bias was always smaller than 6%.

Summarizing, the previous results suggest that ignoring the sampling variability in $\widehat{\mathbf{b}}$ induces a negligible bias in the assessment of the ICA for sample sizes of $N = 100$ patients or larger. Additionally, the relative bias observed when assessing the SPF could be considered generally acceptable for moderate sample sizes ($N \geq 300$), taking values smaller than about 11% in both scenarios. However, there were some substantial differences in the relative bias for the SPF in scenarios 1 and 2, and more simulations may be needed to examine this issue in more detail.

Of course, in the analysis of a real-life case study, the approach detailed in Section 5 can always be used to account for the sampling variability in $\widehat{\mathbf{b}}$.

Table 2: Scenario 1: Relative bias in the estimation of R_H^2 and $r\,(i,\,j)$ as a function of N.

	N				
Parameter	50	100	300	600	1000
R_H^2	0.0855	0.0352	0.0137	0.0044	-0.0032
$r\,(1,\,1)$	0.0254	0.0204	0.0063	0.0027	-0.0011
$r\,(-1,\,1)$	-0.0677	-0.0410	-0.0127	0.0016	0.0089
$r\,(0,\,1)$	0.0125	0.0061	0.0054	0.0025	0.0033
$r\,(1,\,0)$	0.0482	0.0302	0.0084	-0.0017	0.0080
$r\,(-1,\,0)$	0.0038	-0.0002	0.0140	0.0031	-0.0050
$r\,(0,\,0)$	-0.0208	-0.0111	-0.0072	0.0014	0.0010
$r\,(1,\,-1)$	0.0076	0.0153	0.0126	0.0054	0.0023
$r\,(-1,\,-1)$	-0.0419	-0.0337	-0.0196	-0.0055	-0.0012
$r\,(0,\,-1)$	0.0356	0.0251	0.0160	0.0072	0.0055

Table 3: Scenario 2: Relative bias in the estimation of R_H^2 and $r\,(i,\,j)$ as a function of N

	N				
Parameter	50	100	300	600	1000
R_H^2	-0.1479	-0.0239	0.0130	0.0069	0.0047
$r\,(1,\,1)$	0.0297	0.0193	0.0116	0.0047	0.0011
$r\,(-1,\,1)$	-0.5202	-0.3199	-0.1138	-0.0601	-0.0249
$r\,(0,\,1)$	0.0321	0.0208	-0.0115	0.0030	0.0075
$r\,(1,\,0)$	0.0882	0.0384	-0.0154	-0.0199	-0.0198
$r\,(-1,\,0)$	0.0026	-0.0149	-0.0123	-0.0167	-0.0046
$r\,(0,\,0)$	-0.0086	-0.0016	0.0033	0.0042	0.0031
$r\,(1,\,-1)$	-0.3371	-0.2254	-0.1095	-0.0743	-0.0566
$r\,(-1,\,-1)$	-0.0043	0.0011	0.0054	0.0061	0.0070
$r\,(0,\,-1)$	0.1347	0.0726	0.0156	-0.0004	-0.0093

5 Accounting for the sampling variability in the estimates of the marginal probabilities

In the analyses presented in Section 3, the sampling variability in the estimates of the marginal probabilities contained in $\boldsymbol{b}$ was not taken into account. For example, $\pi_{1\cdot 1\cdot}$ was fixed at its estimated value 0.4215 in each run of the `ICA.BinBin.Grid.Sample()` function. To account for the uncertainty in the estimation of $\pi_{1\cdot 1\cdot}$, this parameter can be uniformly sampled from its corresponding confidence interval $CI_{95\%} = [0.3562; 0.4868]$ at each run of the Monte Carlo algorithm and a similar procedure can also be used for the other marginal probabilities. Obviously, the previous sampled components of $\boldsymbol{b}$ are restricted to sum less than one.

The function `ICA.BinBin.Grid.Sample.Uncert()` implements this approach. The following commands can be used to assess the SPF while accounting for the sampling variability in $\widehat{\boldsymbol{b}}$ under the assumption of no monotonicity:

```
ICA_No2 <- ICA.BinBin.Grid.Sample.Uncert(pi1_1_=runif(10000, 0.3562, 0.4868),
pi0_1_=runif(10000, 0.0240, 0.0837), pi1_0_=runif(10000, 0.0240, 0.0837),
pi_1_1=runif(10000, 0.4434, 0.5742), pi_1_0=runif(10000, 0.0081, 0.0533),
pi_0_1=runif(10000, 0.0202, 0.0763), Seed=1, Monotonicity=c("No"), M=10000)

SPF_No2 <- SPF.BinBin(ICA_No2)
```

Notice that the commands used here are similar to those employed in Section 3, being the only difference that the point estimates for the marginals (e.g., `pi1_1_=0.4215`) are now replaced by uniform distributions (e.g., `pi1_1_=runif(10000, 0.3562, 0.4868)`)) in the `ICA.BinBin.Grid.Sample.Uncert()` function call. The fitted objects can subsequently be examined in the same way as was done in Section 3, e.g., the `summary()` and `plot()` functions can be applied to the fitted objects.

Figures 1–4 show the frequency densities for the SPF that were obtained when the sampling variability was (right figures) and was not (left figures) taken into account in the no monotonicity, monotonicity for S, monotonicity for T, and monotonicity for S and T regions. In addition, Tables 4–7 provide the corresponding summary statistics.

Overall, both analyses produced very similar results in the region where monotonicity does not hold (see Figure 1 and Table 4). In the monotonicity for S, T and both S and T regions, both analyses produce a very similar assessment for $r(i, j = 0)$, i.e., in all monotonicity scenarios $r(0, 0)$ tended to be high and $r(1, 0)$, $r(-1, 0)$ tended to be low. Thus, a lack of effect on BPRS ($\Delta S = 0$) seems to be indicative of a lack of effect on PANSS ($\Delta T = 0$) in all monotonicity scenarios, whether sampling variability is taken into account or not. However, the assessment of $r(i, j = 1)$ differed in both analysis. In general, when the sampling variability was accounted for, the measures of central tendency were larger for $r(1, 1)$ and smaller for $r(-1, 1)$, $r(0, 1)$.

In the monotonicity for T region, the assessment of $r(i, j = -1)$ also seems to differ in both analysis. For example, the mean $r(0, -1) = 0.7587$ and $r(1, -1) = 0.2413$ when sampling variability is not accounted for, and the mean $r(0, -1) = 0.9345$ and $r(1, -1) = 0.0655$ when sampling variability is accounted for.

	Sampling variability marginals not accounted for					Sampling variability marginals accounted for				
	Mean	Median	Mode	*SD*	[min; max]	Mean	Median	Mode	*SD*	[min; max]
$r(-1, -1)$	0.7484	0.8251	0.8711	0.2172	[0.0585; 0.9717]	0.8050	0.8359	0.8648	0.1347	[0.0087; 0.9930]
$r(0, -1)$	0.2033	0.1357	0.1034	0.1846	[0.0047; 0.7679]	0.1748	0.1463	0.1138	0.1272	[0.0033; 0.9232]
$r(1, -1)$	0.0483	0.0235	0.0150	0.0832	[0.0001; 0.6067]	0.0202	0.0161	0.0063	0.0190	[0.0001; 0.2134]
$r(-1, 0)$	0.0781	0.0582	0.0421	0.0561	[0.0060; 0.2521]	0.1077	0.0907	0.0545	0.0753	[0.0008; 0.7185]
$r(0, 0)$	0.8597	0.8906	0.9153	0.0874	[0.5476; 0.9705]	0.8043	0.8357	0.8733	0.1155	[0.0599; 0.9930]
$r(1, 0)$	0.0622	0.0473	0.0395	0.0481	[0.0037; 0.3107]	0.0880	0.0736	0.0586	0.0606	[0.0023; 0.5438]
$r(-1, 1)$	0.0443	0.0310	0.0224	0.0454	[0.0007; 0.2324]	0.0312	0.0274	0.0093	0.0230	[0.0001; 0.1593]
$r(0, 1)$	0.1303	0.1118	0.0790	0.0896	[0.0044; 0.4287]	0.1261	0.1108	0.0920	0.0673	[0.0039; 0.5064]
$r(1, 1)$	0.8254	0.8473	0.8838	0.0998	[0.4556; 0.9740]	0.8472	0.8585	0.8878	0.0728	[0.4715; 0.9939]

Table 4: SPF summary statistics under the no monotonicity assumption when the sampling variability in the marginal probabilities is not accounted for (left) and is accounted for (right)

	Sampling variability marginals not accounted for					Sampling variability marginals accounted for				
	Mean	Median	Mode	*SD*	[min; max]	Mean	Median	Mode	*SD*	[min; max]
$r(-1,\ 0)$	0.0281	0.0249	0.0173	0.0167	[0.0028; 0.0670]	0.0768	0.0775	0.0865	0.0347	[0.0249; 0.1486]
$r(0,\ 0)$	0.9091	0.9074	0.9045	0.0246	[0.8456; 0.9715]	0.8774	0.8832	0.8928	0.0455	[0.7707; 0.9361]
$r(1,\ 0)$	0.0629	0.0612	0.0589	0.0151	[0.0201; 0.0899]	0.0459	0.0444	0.0402	0.0177	[0.0139; 0.0807]
$r(-1,\ 1)$	0.1213	0.1172	0.0512	0.0907	[0.0005; 0.4446]	0.0357	0.0333	0.0360	0.0288	[0.0005; 0.1051]
$r(0,\ 1)$	0.3633	0.3408	0.3255	0.1789	[0.0520; 0.7667]	0.1444	0.1442	0.1039	0.0702	[0.0326; 0.2898]
$r(1,\ 1)$	0.5153	0.50723	0.4783	0.1838	[0.1018; 0.9247]	0.8199	0.8030	0.7632	0.0726	[0.7034; 0.9541]

Table 5: SPF summary statistics under the monotonicity for S assumption when the sampling variability in the marginal probabilities is not accounted for (left) and is accounted for (right)

	Sampling variability marginals not accounted for					Sampling variability marginals accounted for				
	Mean	Median	Mode	*SD*	[min; max]	Mean	Median	Mode	*SD*	[min; max]
$r(0, -1)$	0.7587	0.7740	0.8970	0.1776	[0.0876; 0.9997]	0.9345	0.9535	0.9736	0.0664	[0.6936; 0.9989]
$r(1, -1)$	0.2413	0.2260	0.1030	0.1776	[0.0003; 0.9124]	0.0655	0.0465	0.0264	0.0664	[0.0011; 0.3064]
$r(0, 0)$	0.9741	0.9767	0.9800	0.0154	[0.9377; 0.9977]	0.9708	0.9727	0.9747	0.0131	[0.9292; 0.9978]
$r(1, 0)$	0.0259	0.0233	0.0200	0.0154	[0.0023; 0.0623]	0.0292	0.0273	0.0254	0.0131	[0.0022; 0.0708]
$r(0, 1)$	0.6992	0.6848	0.6641	0.1083	[0.4541; 0.9390]	0.4638	0.4390	0.4275	0.1639	[0.1751; 0.8578]
$r(1, 1)$	0.3008	0.3152	0.3359	0.1083	[0.0610; 0.5459]	0.5362	0.5611	0.5726	0.1639	[0.1422; 0.8249]

Table 6: SPF summary statistics under the monotonicity for T assumption when the sampling variability in the marginal probabilities is not accounted for (left) and is accounted for (right)

	Sampling variability marginals not accounted for					Sampling variability marginals accounted for				
	Mean	Median	Mode	*SD*	[min; max]	Mean	Median	Mode	*SD*	[min; max]
$r(0,\,0)$	0.9556	0.9545	0.9486	0.0148	[0.9301; 0.9930]	0.9552	0.9554	0.9621	0.0174	[0.9083; 0.9983]
$r(1,\,0)$	0.0444	0.0455	0.0514	0.0148	[0.0070; 0.0700]	0.0448	0.0446	0.0380	0.0174	[0.0017; 0.0917]
$r(0,\,1)$	0.7136	0.7255	0.7915	0.1667	[0.2928; 0.9999]	0.5873	0.5748	0.4988	0.2043	[0.1039; 0.9999]
$r(1,\,1)$	0.2864	0.2745	0.2086	0.1667	[0.0001; 0.7072]	0.4127	0.4253	0.5012	0.2043	[0.0001; 0.8962]

Table 7: SPF summary statistics under the monotonicity for S and T assumption when the sampling variability in the marginal probabilities is not accounted for (left) and is accounted for (right)

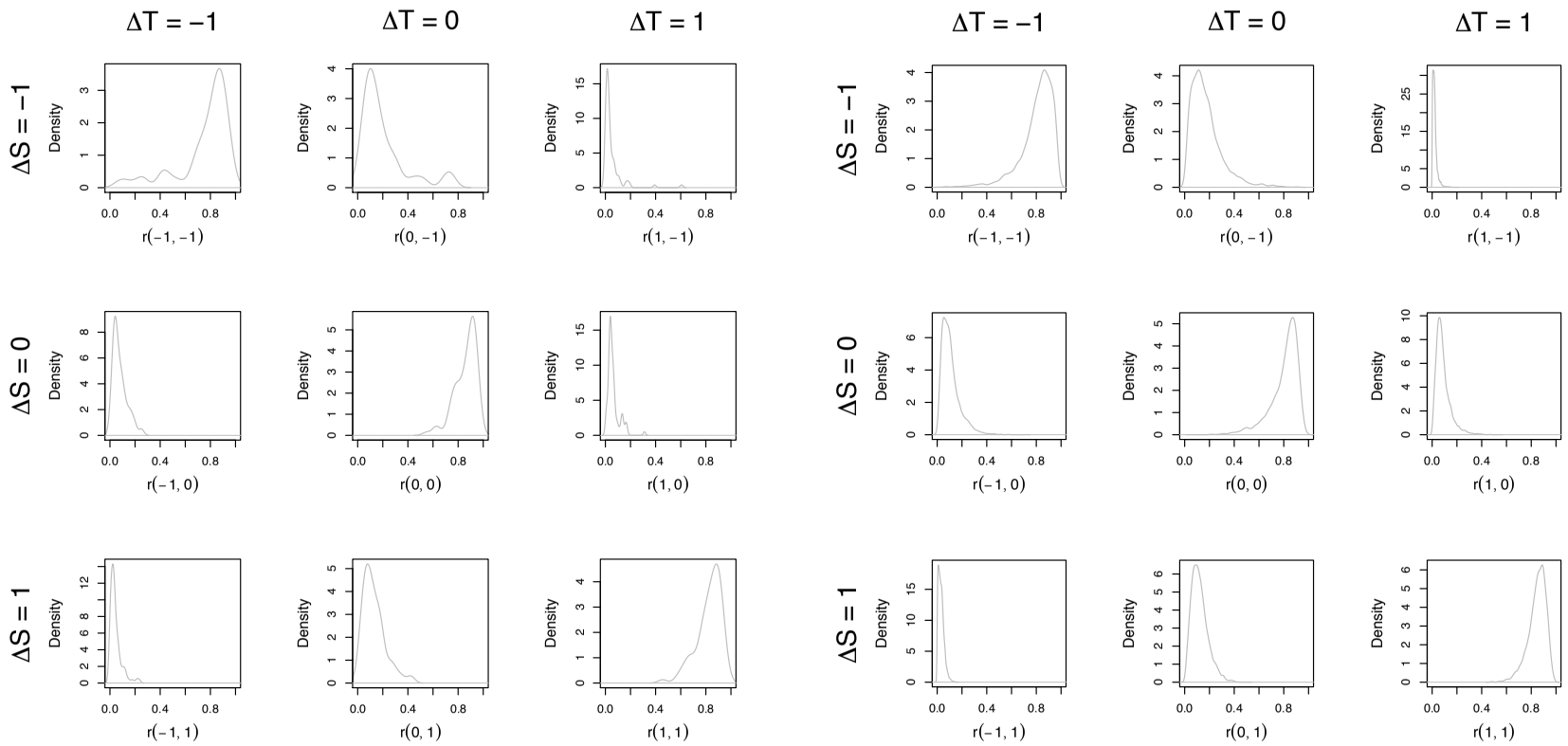


Figure 1: SPF densities under the no monotonicity assumption when the sampling variability in the marginal probabilities is not accounted for (left plot) and is accounted for (right plot)

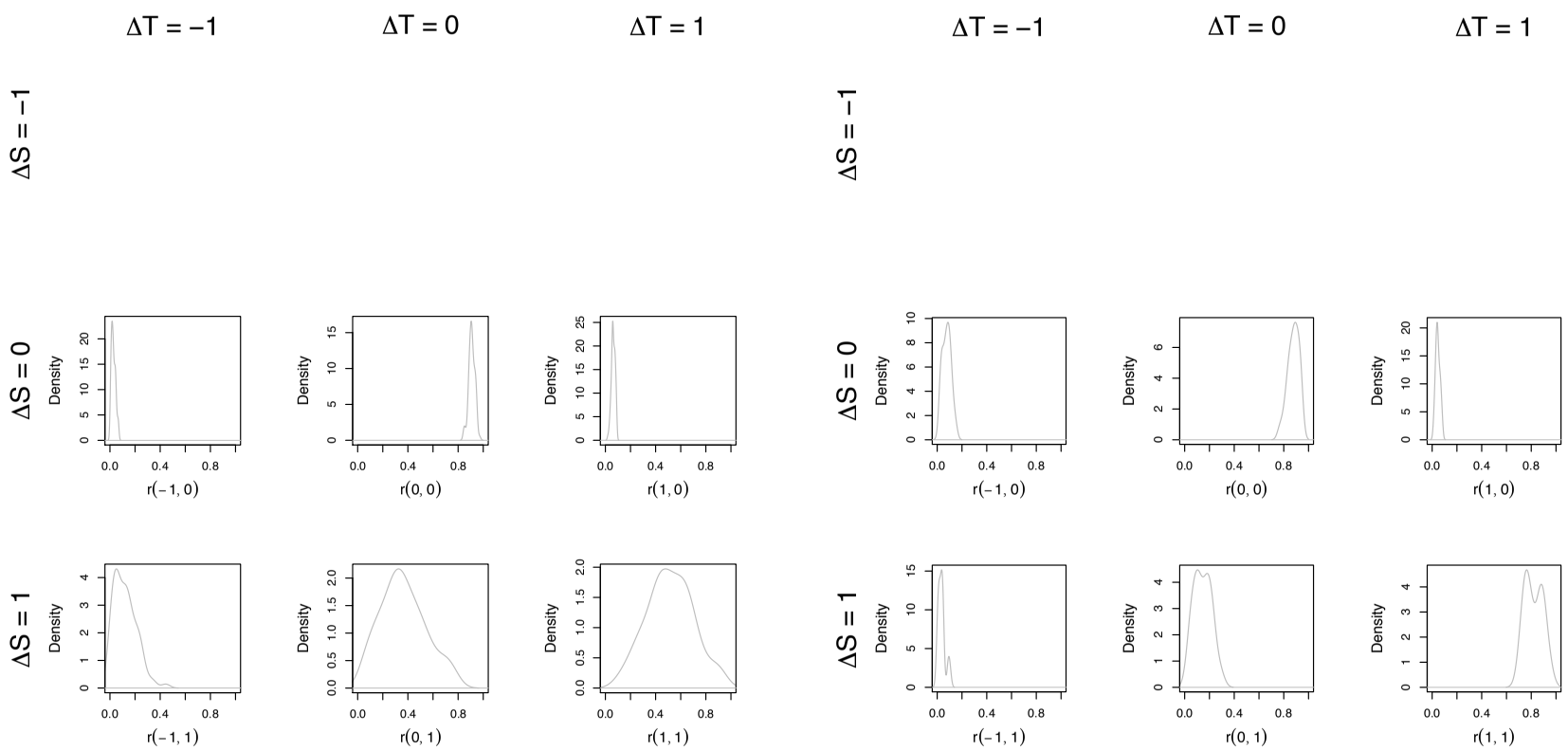


Figure 2: SPF densities under the monotonicity for S assumption when the sampling variability in the marginal probabilities is not accounted for (left plot) and is accounted for (right plot)

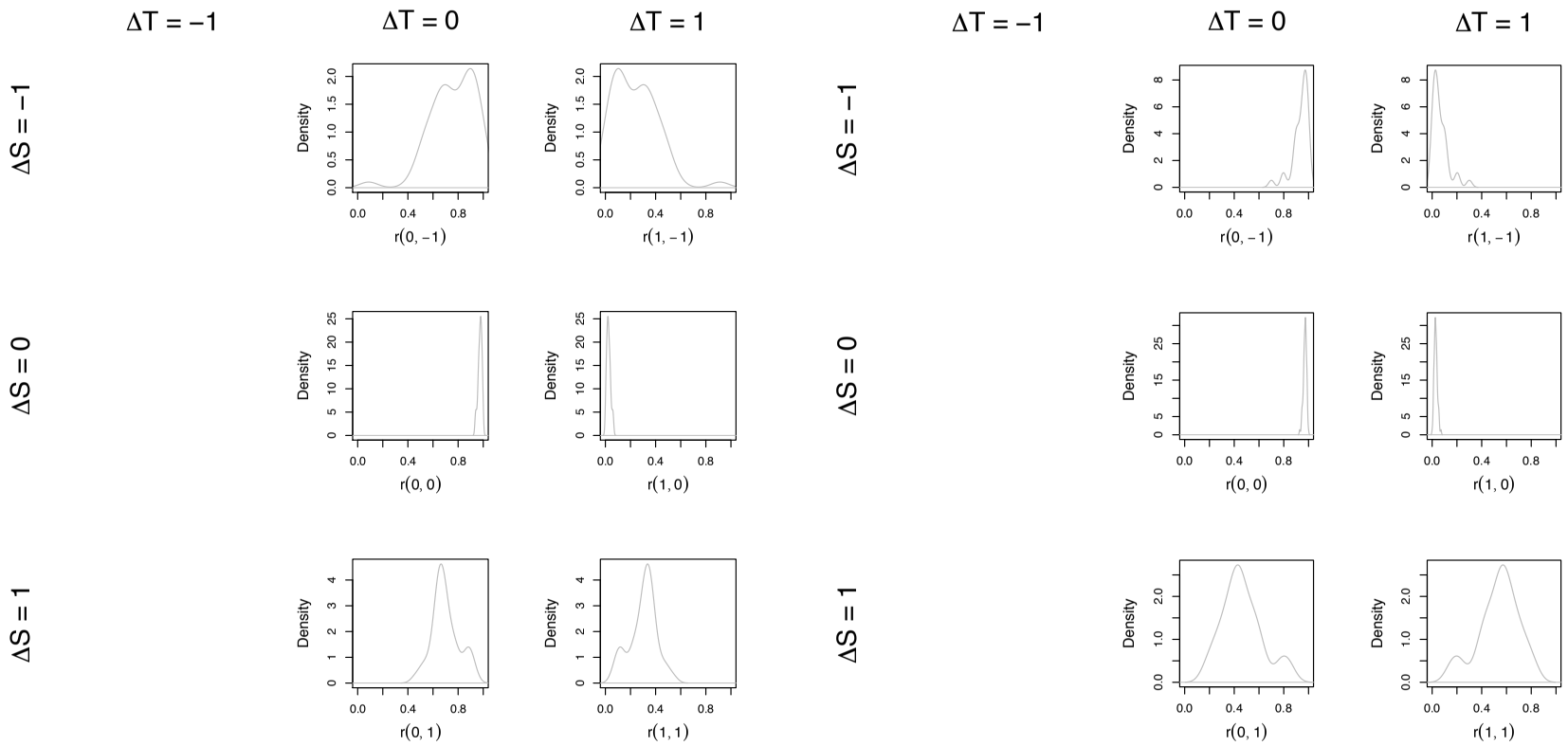


Figure 3: SPF densities under the monotonicity for T assumption when the sampling variability in the marginal probabilities is not accounted for (left plot) and is accounted for (right plot)

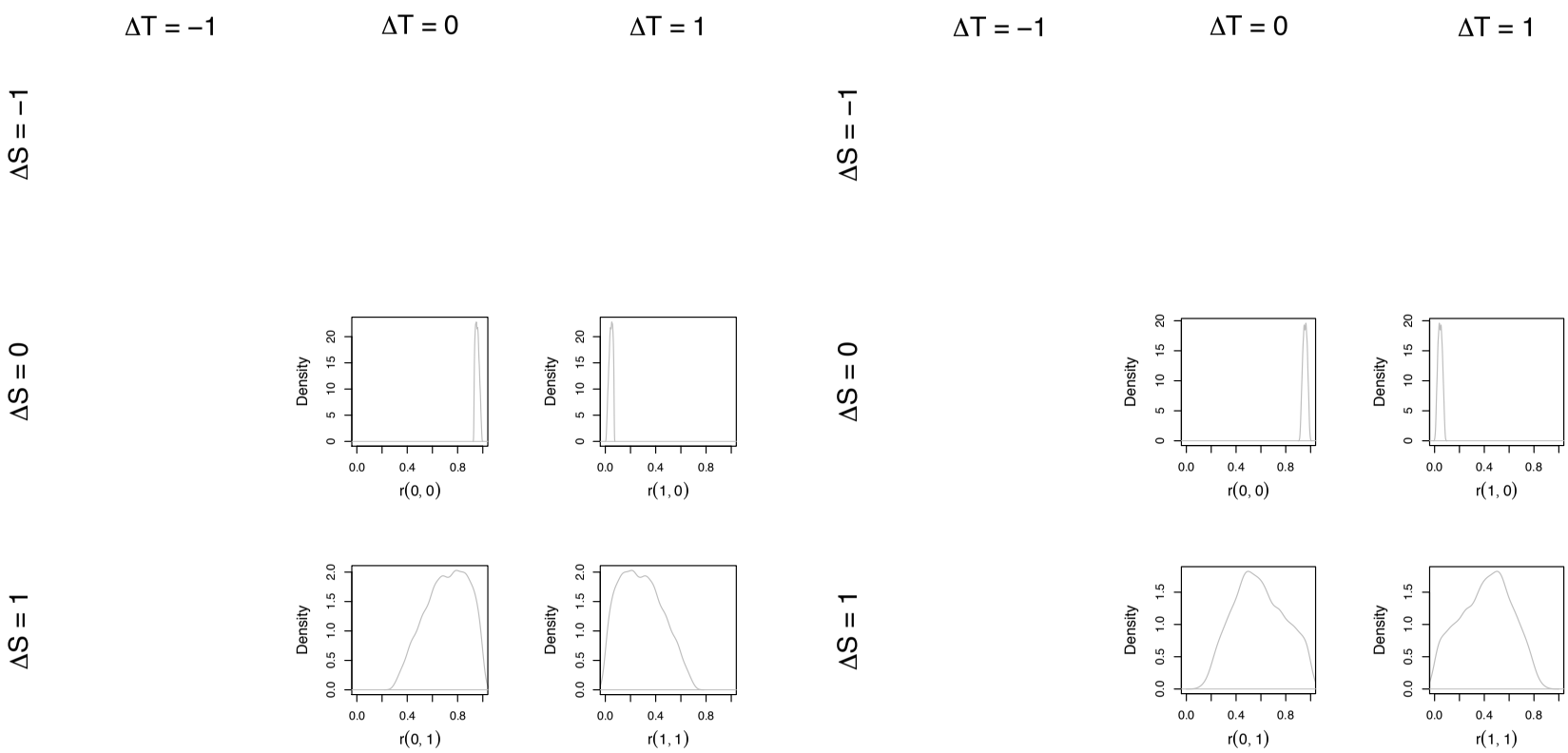


Figure 4: SPF densities under the monotonicity for S and T assumption when the sampling variability in the marginal probabilities is not accounted for (left plot) and is accounted for (right plot)

6 Algebraic developments and definitions

Geometrically characterizing Γ_D

The distribution of $\mathbf{Y}$ can be tabulated as in Table 8 and the set of restrictions on $\boldsymbol{\pi}$ can be written as

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

with the points in the sub-indexes indicating sums over those specific sub-indexes. Further, if one defines the vector $\boldsymbol{b}' = (1, \pi_{1\cdot 1\cdot}, \pi_{1\cdot 0\cdot}, \pi_{\cdot 1\cdot 1}, \pi_{\cdot 1\cdot 0}, \pi_{0\cdot 1\cdot}, \pi_{\cdot 0\cdot 1})$, and the matrix

$$\mathbf{A} = \begin{pmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1\\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1\\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0\\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1\\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0\\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0
\end{pmatrix},$$

then all the identified restrictions given in (1) can be written as a system of linear equations,

$$\mathbf{A}\boldsymbol{\pi} = \boldsymbol{b}, \tag{2}$$

where the vector of parameters $\boldsymbol{\pi}$ is ordered as in Table 8. The hyperplane given in (2) geometrically characterizes the subspace of Γ compatible with the data at hand, i.e., $\Gamma_D = \{\boldsymbol{\pi} \in \Gamma : \mathbf{A}\boldsymbol{\pi} = \boldsymbol{b}\}$. The matrix $\mathbf{A}$ has rank 7 and can be partitioned as $\mathbf{A} = \left(\mathbf{A}_r | \mathbf{A}_f\right)$ where $\mathbf{A}_f$ denotes the submatrix given by the last 9 columns of $\mathbf{A}$ and $\mathbf{A}_r$ is a full column rank matrix. Similarly, the vector $\boldsymbol{\pi}$ can be partitioned as $\boldsymbol{\pi}' = \left(\boldsymbol{\pi}_r' | \boldsymbol{\pi}_f'\right)$ with $\boldsymbol{\pi}_f$ the subvector given by the last 9 components of $\boldsymbol{\pi}$. Using these partitions (2) can be rewritten as $\mathbf{A}_r\boldsymbol{\pi}_r + \mathbf{A}_f\boldsymbol{\pi}_f = \boldsymbol{b}$.

Proof of Lemma 2. Let us consider a general function $\psi : \{-1,0,1\} \to \{-1,0,1\}$, we want to find the function ψ_b that maximizes the probability

$$\begin{aligned}
P\left[\Delta T = \psi(\Delta S)\right] &= \sum_{i=-1}^{1} P\left(\Delta T = i, \psi(\Delta S) = i\right), \\
&= \sum_{i=-1}^{1} \sum_{j \in \psi^{-1}(i)} P\left(\Delta T = i, \Delta S = j\right), \\
&= \sum_{i=-1}^{1} \sum_{j \in \psi^{-1}(i)} P\left(\Delta T = i | \Delta S = j\right) P\left(\Delta S = j\right),
\end{aligned} \tag{3}$$

where $\psi^{-1}(i) = \{j \in \{-1,0,1\} : \psi(j) = i\}$. If $P\left(\Delta S = j\right) = 0$ then $P\left(\Delta T = i | \Delta S = j\right)$ does not contribute to the probability $P\left[\Delta T = \psi(\Delta S)\right]$ and, hence, one can focus only on the support of ΔS. Without loss of generality, let us assume that $P\left(\Delta S = j\right) > 0$ for all j. Basically, ψ can be thought of as a correspondence between the column numbers and the row numbers of the distribution of $\boldsymbol{\Delta}$, i.e., between the column and row numbers of table 1 in the manuscript, where every column number j gets mapped into one and only one row number i. Therefore, defining a function ψ is equivalent to choosing 3 cells in table 1, each of them located in a different column. Consequently, maximizing $P\left[\Delta T = \psi(\Delta S)\right]$ is

Table 8: Distribution of $\mathbf{Y}$

$\boldsymbol{\pi}$	T_0	T_1	S_0	S_1
π_{0000}	0	0	0	0
π_{0100}	0	1	0	0
π_{0010}	0	0	1	0
π_{0001}	0	0	0	1
π_{0101}	0	1	0	1
π_{1000}	1	0	0	0
π_{1010}	1	0	1	0
π_{1001}	1	0	0	1
π_{1110}	1	1	1	0
π_{1101}	1	1	0	1
π_{1011}	1	0	1	1
π_{1111}	1	1	1	1
π_{0110}	0	1	1	0
π_{0011}	0	0	1	1
π_{0111}	0	1	1	1
π_{1100}	1	1	0	0

equivalent to maximizing the sum of the corresponding cells probabilities in each column or, in other words, maximizing $P(\Delta T = i|\Delta S = j)$ for $j = -1, 0, 1$ and, thus,

$$\psi_b(j) = \arg\max_i r(i,j) = \arg\max_i P(\Delta T = i|\Delta S = j)$$

□

Associative and dissociative proportions

Frangakis and Rubin (2002) 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* (AE) relative to its *dissociative effect* (DE). 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 (Li, Taylor and Elliott, 2010; Elliott, Li and Taylor, 2013).

Using the notation in Table 1 in Alonso *et al.* (2016b), the definitions given in Elliott, Li and Taylor (2013) take the form: $AE = \left(\pi^{\Delta}_{11} - \pi^{\Delta}_{-11}\right) + \left(\pi^{\Delta}_{1-1} - \pi^{\Delta}_{-1-1}\right)$, i.e., AE is the net treatment effect on patients whose surrogate was responsive to the treatment. Furthermore, $DE = \pi^{\Delta}_{10} - \pi^{\Delta}_{-10}$, i.e., DE is the net treatment effect on patients whose surrogate was not responsive to the treatment. Finally, the causal treatment effect on the true endpoint is defined as $CET = \pi^{\Delta T}_{1} - \pi^{\Delta T}_{-1}$, i.e., the net treatment effect corresponding to the fraction responsive to the treatment minus the fraction harmed. Because AE and DE are constrained to sum CET, Taylor, Wang and Thıébaut (2005) proposed to use instead the so-called associative ($AP = AC/CET$) and dissociative ($DP = DE/CET$) proportions respectively. A good surrogate is then expected to have a large AP and a small DP. Using some theoretical elements Alonso *et al.* (2016a) showed that, at least in some scenarios, the ICA may offer a more coherent assessment of surrogacy than the AP and DP.

References

Alonso A, Van der Elst W, Molenberghs G, Buyse M and Burzykowski T. (2015). A causal-inference approach for the validation of surrogate endpoints based on information theory and sensitivity analysis *Biometrics*. (accepted)

Alonso, A., and Van der Elst, W. (2015). Assessing a surrogate effect predictive value in a causal-inference framework. *Submitted*.

Buyse, M., Burzykowski, T., Alonso, A., and Molenberghs, G. (2014). Direct estimation of joint counterfactual probabilities, with application to surrogate marker validation. *Submitted*.

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

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

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

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.

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

Prentice, R.L. (1989). Surrogate endpoints in clinical trials: definitions and operational criteria. *Statistics in Medicine*, **8**, 431–440.

Singh, M., and 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.

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

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.