Research Article

Statistics in Medicine

Received 30 May 2014, Accepted 23 October 2015 Published online 26 November 2015 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6807

# Exploring the relationship between the causal-inference and meta-analytic paradigms for the evaluation of surrogate endpoints

Wim Van der Elst,[a*†] Geert Molenberghs[a,b] and Ariel Alonso[b]

**Nowadays, two main frameworks for the evaluation of surrogate endpoints, based on causal-inference and meta-analysis, dominate the scene. Earlier work showed that the metrics of surrogacy introduced in both paradigms are related, although in a complex way that is difficult to study analytically. In the present work, this relationship is further examined using simulations and the analysis of a case study. The results indicate that the extent to which both paradigms lead to similar conclusions regarding the validity of the surrogate, depends on a complex interplay between multiple factors like the ratio of the between and within trial variability and the unidentifiable correlations between the potential outcomes. All the analyses were carried out using the newly developed R package *Surrogate*, which is freely available via CRAN. Copyright © 2015 John Wiley & Sons, Ltd.**

**Keywords:** causal-inference approach; meta-analytic approach; surrogate markers; R package surrogate

## 1. Introduction

Often the duration, complexity and cost of a clinical trial are substantially affected by the endpoint used to assess treatment efficacy [1]. Actually, in some situations, the most credible indicator of therapeutic response, the so-called true endpoint, may be distant in time (e.g., survival time in early cancer stages), rare (e.g., pregnancy in severe luteinizing hormone deficiency), ethically challenging (e.g., procedures that involve a non-negligible health risk), or expensive (e.g., imaging data). An appealing strategy in these circumstances is to substitute the true endpoint by a so-called surrogate endpoint that can be measured earlier, occurs more frequently, is more ethically acceptable or cheaper. If such an endpoint further allows for a precise prediction of the clinical effect of the treatment on the true endpoint, then it is termed a valid surrogate endpoint [2–7].

The statistical evaluation of surrogate endpoints is not a trivial endeavor, and over the last decades various strategies have been developed for this purpose. Most of these methods can be classified broadly along two dimensions, taking into account (1) whether they use information from a single or from multiple clinical trials and (2) whether they focus on individual or expected causal effects to carry out the validation exercise [1, 2, 8–10]. Several authors have argued that there is a relationship between the metrics of surrogacy that are used in the individual and expected causal effect paradigms [11, 12], but the complex nature of this relationship hinders the use of analytical techniques to gain a deeper understanding. One way to deal with this issue is to simplify the problem by making additional assumptions. For example, when it can be assumed that the potential outcomes for the surrogate and the true endpoints are independent, it can be heuristically shown that a surrogate that is successfully evaluated in the meta-analytic framework will typically also be appealing for proponents of the causal-inference framework [12]. However, the latter assumption is generally implausible. Indeed, it is natural to expect that the potential outcomes for the surrogate and the true endpoints within the same patients are correlated, because of,

[a] *I-BioStat, Universiteit Hasselt, B-3590 Diepenbeek, Belgium*
[b] *I-BioStat, Katholieke Universiteit Leuven, B-3000 Leuven, Belgium*
**Correspondence to: Wim Van der Elst, I-BioStat, Universiteit Hasselt, Agoralaan 1, 3590 Diepenbeek, Belgium.*
[†] *E-mail: wim.Vanderelst@gmail.com*

for example, genetic or health factors. The first aim of the present study was to evaluate the relationship between the meta-analytic and causal-inference based metrics of surrogacy in more realistic scenarios (i.e., under less restrictive assumptions) using simulations. Particular interest will be in the identification of the conditions under which the frameworks based on individual and expected causal effects lead to similar conclusions regarding the validity of the putative surrogate.

A problem that analysts often face when evaluating surrogate endpoints is the lack of user-friendly software packages to conduct the complex analyses. The second aim of the current work is to present the R package *Surrogate* (freely available at CRAN) and exemplify its use in the analysis of real-life data. The package allows for the evaluation of surrogate endpoints based on individual and expected causal effects in the single-trial and multiple-trial settings. Having such a tool available is useful for practitioners, that is, it can be examined in a straightforward way whether the different surrogate evaluation paradigms lead to the same general conclusion regarding the appropriateness of the putative surrogate.

The remainder of this paper is organized as follows. In Sections 2–5, the theoretical models that underlie the causal-inference and the meta-analytic frameworks in the single-trial and multiple-trial settings are detailed, and the results of the simulation studies are described. In Section 6, the implementation of the single-trial and multiple-trial meta-analytic and causal-inference approaches is illustrated using the data of a clinical trial in schizophrenia. In Section 7, some final comments and recommendations for future studies are provided.

## 2. Single-trial setting

We will start by considering the so-called single-trial setting (STS), that is, the setting in which information on both the surrogate and true endpoints is available from a single clinical trial. In this setting, the inclusion and exclusion criteria of the trial characterize a unique, well-defined, and fixed population in which the validation exercise is framed. Further, it will be assumed that merely two treatments are under evaluation ($Z = 0/1$).

### 2.1. *Individual causal association*

Following Rubin's causal inference model [13], it will be assumed that each patient $j$ has four potential outcomes: $T_{0j}$, $T_{1j}$, $S_{0j}$, and $S_{1j}$, denoting the outcomes for the true ($T$) and surrogate ($S$) endpoints under the control ($Z = 0$) and the experimental ($Z = 1$) conditions, respectively. The individual causal effects of $Z$ on $T$ and $S$ for a given subject $j$ can then be defined as $\Delta_{T_j} = T_{1j} - T_{0j}$ and $\Delta_{S_j} = S_{1j} - S_{0j}$, and the expected causal effects in the population of interest as $\beta = E(T_{1j} - T_{0j})$ and $\alpha = E(S_{1j} - S_{0j})$.

A fundamental problem in the causal-inference framework is that $\Delta_{T_j}$ and $\Delta_{S_j}$ are not identifiable from the data, basically, because only two out of the four potential outcomes are observable [14]. In fact, if $T_j$ and $S_j$ denote the observed responses for subject $j$, then under the Stable Unit Treatment Value Assumption (SUTVA), one has that $T_j = Z_jT_{1j} + (1 - Z_j)T_{0j}$ and $S_j = Z_jS_{1j} + (1 - Z_j)S_{0j}$. Under the additional assumptions of independence $Z_j \perp (T_{0j}, T_{1j})$ and $Z_j \perp (S_{0j}, S_{1j})$, it further holds that $\beta = E(T_j \mid Z_j = 1) - E(T_j \mid Z_j = 0)$ and $\alpha = E(S_j \mid Z_j = 1) - E(S_j \mid Z_j = 0)$. The latter quantities are estimable based on the means of $T$ and $S$ in the experimental and the control groups and, therefore, the expected causal effects are identifiable under rather general conditions. Notice that the independence between the potential outcomes and the treatment variable is guaranteed in randomized clinical trials because of the random treatment allocation.

In the rest of the manuscript, it will be assumed that the vector of potential outcomes $\boldsymbol{Y}_j = (T_{0j}, T_{1j}, S_{0j}, S_{1j})'$ has distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (\mu_{T_0}, \mu_{T_1}, \mu_{S_0}, \mu_{S_1})'$ and

$$\boldsymbol{\Sigma} = \left(\begin{array}{cc|cc} \sigma_{T_0T_0} & \sigma_{T_0T_1} & \sigma_{T_0S_0} & \sigma_{T_0S_1} \\ \sigma_{T_0T_1} & \sigma_{T_1T_1} & \sigma_{T_1S_0} & \sigma_{T_1S_1} \\ \hline \sigma_{T_0S_0} & \sigma_{T_1S_0} & \sigma_{S_0S_0} & \sigma_{S_0S_1} \\ \sigma_{T_0S_1} & \sigma_{T_1S_1} & \sigma_{S_0S_1} & \sigma_{S_1S_1} \end{array}\right)$$

The previous distributional assumption implies

$$\boldsymbol{\Delta}_j = \mathbf{A}\boldsymbol{Y}_j = \begin{pmatrix} T_{1j} - T_{0j} \\ S_{1j} - S_{0j} \end{pmatrix} \sim N\left(\boldsymbol{\mu}_{\Delta}, \boldsymbol{\Sigma}_{\Delta}\right), \qquad (1)$$

where $\mathbf{\Sigma}_\Delta = \mathbf{A}\mathbf{\Sigma}\mathbf{A}'$, $\boldsymbol{\mu}_\Delta = (\beta, \alpha)'$ and $\mathbf{A}$ is the corresponding contrast matrix. Alonso *et al.* [12] argued in favor of assessing surrogacy in the STS using the so-called individual causal association (ICA) $\rho_\Delta = \text{corr}\left(\Delta_{Tj}, \Delta_{Sj}\right)$ and showed that

$$\rho_\Delta = \frac{\sqrt{\sigma_{T_0T_0}\sigma_{S_0S_0}}\rho_{T_0S_0} + \sqrt{\sigma_{T_1T_1}\sigma_{S_1S_1}}\rho_{T_1S_1} - \sqrt{\sigma_{T_1T_1}\sigma_{S_0S_0}}\rho_{T_1S_0} - \sqrt{\sigma_{T_0T_0}\sigma_{S_1S_1}}\rho_{T_0S_1}}{\sqrt{\left(\sigma_{T_0T_0} + \sigma_{T_1T_1} - 2\sqrt{\sigma_{T_0T_0}\sigma_{T_1T_1}}\rho_{T_0T_1}\right)\left(\sigma_{S_0S_0} + \sigma_{S_1S_1} - 2\sqrt{\sigma_{S_0S_0}\sigma_{S_1S_1}}\rho_{S_0S_1}\right)}}, \tag{2}$$

where $\rho_{XY}$ denotes the correlation between the potential outcomes $X$ and $Y$. Notice that ICA is also a measure of prediction accuracy, that is, 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. If one further assumes that $\sigma_{T_0T_0} = \sigma_{T_1T_1} = \sigma_T$ and $\sigma_{S_0S_0} = \sigma_{S_1S_1} = \sigma_S$, that is, the variability of the true and the surrogate endpoints is constant across the two treatment conditions, then expression (2) simplifies to

$$\rho_\Delta = \frac{\rho_{S_0T_0} + \rho_{S_1T_1} - \rho_{S_0T_1} - \rho_{S_1T_0}}{2\sqrt{\left(1 - \rho_{T_0T_1}\right)\left(1 - \rho_{S_0S_1}\right)}}. \tag{3}$$

The previous assumption of homoscedasticity is used in many statistical models, and it is testable. For simplicity, homoscedasticity will be assumed in the remainder of Section 2 and in Sections 3–5, but the derived conclusions are valid in the more general setting as well (Section 6).

### 2.2. *Individual causal association: some identifiability issues*

The practical use of ICA is challenging. Indeed, the correlations $\rho_{S_0T_1}$, $\rho_{S_1T_0}$, $\rho_{T_0T_1}$, and $\rho_{S_0S_1}$ in (3) are not estimable from the data [9, 10, 12], and, consequently, $\rho_\Delta$ is not identifiable. Two strategies are possible to deal with these identifiability issues. First, one can try to define plausible identifiability conditions based on biological or subject-specific knowledge. However, such subject-specific knowledge may not always be available and/or these biologically plausible assumptions often have to be supplemented with additional assumptions for which no such subject-specific knowledge exists. In addition, different identifiability conditions can lead to substantially different estimates of $\rho_\Delta$ and thus to different conclusions regarding the appropriateness of the surrogate.

A second approach is to implement a simulation-based sensitivity analysis in which $\rho_\Delta$ is estimated across a set of plausible values for the unidentifiable correlations. Essentially, in a first step, grids of values $G = \left\{g_1, g_2, \ldots, g_k\right\}$ are considered for the unidentified correlations between the potential outcomes. Next, several $\Sigma$ matrices are generated fixing the identifiable correlations $\rho_{S_0T_0}$, $\rho_{S_1T_1}$ at their estimated values and considering all the combinations emanating from the specified grids for the unidentified correlations $\rho_{S_0T_1}$, $\rho_{S_1T_0}$, $\rho_{T_0T_1}$, and $\rho_{S_0S_1}$. From all the previous $\Sigma$ matrices, only those that are positive definite (i.e., valid correlation matrices) are used in the subsequent step. Finally, $\rho_\Delta$ is estimated based on these positive definite matrices. Intuitively, the so-obtained vector $\rho_\Delta$ quantifies the individual causal association across all plausible 'worlds', that is, across those scenarios where the assumptions made for the unidentified correlations are compatible with the observed data ($\widehat{\rho}_{S_0T_0}$ and $\widehat{\rho}_{S_1T_1}$). The general behavior of $\rho_\Delta$ can subsequently be examined, for example, by quantifying the variability and the range of its estimates, and in this way the sensitivity of the results with respect to the unverifiable assumptions can be assessed.

Importantly, these two strategies to deal with the identifiability issues are not mutually exclusive. In fact, the simulation-based approach previously described allows for a straightforward incorporation of subject-specific knowledge in case it is available. For example, if it is reasonable to assume, based on biological knowledge, that a particular unidentified correlation is positive, then a grid $G$ that only contains positive values can be used for this correlation when carrying out the sensitivity analysis.

### 2.3. *Expected causal association*

Although conceptually attractive, measures based on individual causal effects pose huge challenges for the evaluation of surrogate endpoints. For example, as previously stated, ICA is not identifiable from the data and cannot be estimated without imposing untestable restrictions on the association structure of the potential outcomes. An alternative approach is to carry out the validation exercise based on the expected causal effects. Alonso *et al.* [12] argued that there may be some practical and methodological reasons to justify this choice. Basically, the expected causal effects: (1) have a causal interpretation; (2) are identifiable under randomization; and (3) may be more appealing to regulatory authorities and, hence, more useful for drug approval. Therefore, one may try to establish surrogacy by studying the expected causal association (ECA), that is, the association between the expected causal effects of the treatment on the surrogate and true endpoint.

Along these lines, Buyse and Molenberghs [3] introduced two quantities to assess surrogacy in the STS: the relative effect (*RE*) and the adjusted association ($\gamma$). These metrics are defined as $RE = \beta/\alpha$ and $\gamma = \text{corr}(T_j, S_j \mid Z_j)$, respectively. The adjusted association quantifies the accuracy by which $T_j$ can be predicted based on $S_j$ in an individual patient, taking treatment into account. The relative effect moves the validation process away from the unidentifiable individual causal effects to the identifiable expected causal effects. However, *RE* only provides information about ECA under strong and unverifiable assumptions. The fundamental problem is that, in the STS, only a single observation, namely, the vector of treatment effects $(\alpha, \beta)$ is available for the estimation of ECA.

Unlike $\rho_\Delta$, both *RE* and $\gamma$ are estimable from the data and, under the (testable) assumption $\rho_{S_0T_0} = \rho_{S_1T_1} = \gamma$, it was shown that [12]

$$|\rho_\Delta - a\gamma| \leqslant b\sqrt{\left(1-\gamma^2\right)}, \tag{4}$$

where $a = \sqrt{\dfrac{1-\rho_{S_0S_1}}{1-\rho_{T_0T_1}}}$ and $b = \sqrt{\dfrac{1+\rho_{S_0S_1}}{1-\rho_{T_0T_1}}}$. Expression (4) clearly shows that $\rho_\Delta$ and $\gamma$ are intrinsically related and that a strong positive correlation between the surrogate and the true endpoint may be considered an indication of a possible positive and large individual causal association between them. Actually, the function $a\gamma$ can be interpreted as an approximation of ICA, with the approximation improving as the adjusted association increases. Unfortunately, the relationship between ICA and the adjusted association is largely determined by the correlations between the potential outcomes for the surrogate ($\rho_{S_0S_1}$) and true endpoint ($\rho_{T_0T_1}$). Both $\rho_{S_0S_1}$ and $\rho_{T_0T_1}$ are unidentifiable from the data and, although theoretically valuable, (4) does not allow to define a threshold for the adjusted association $\gamma$ that guarantees a large positive value for $\rho_\Delta$ in all scenarios.

In the next section, the relationship between these important concepts will be further explored via simulation. The idea is to clarify, for example, how sensitive $\rho_\Delta$ is with respect to the assumptions regarding the unidentified correlations or how large $\gamma$ should be in order to likely produce a large $\rho_\Delta$.

## 3. Single-trial setting: simulation study

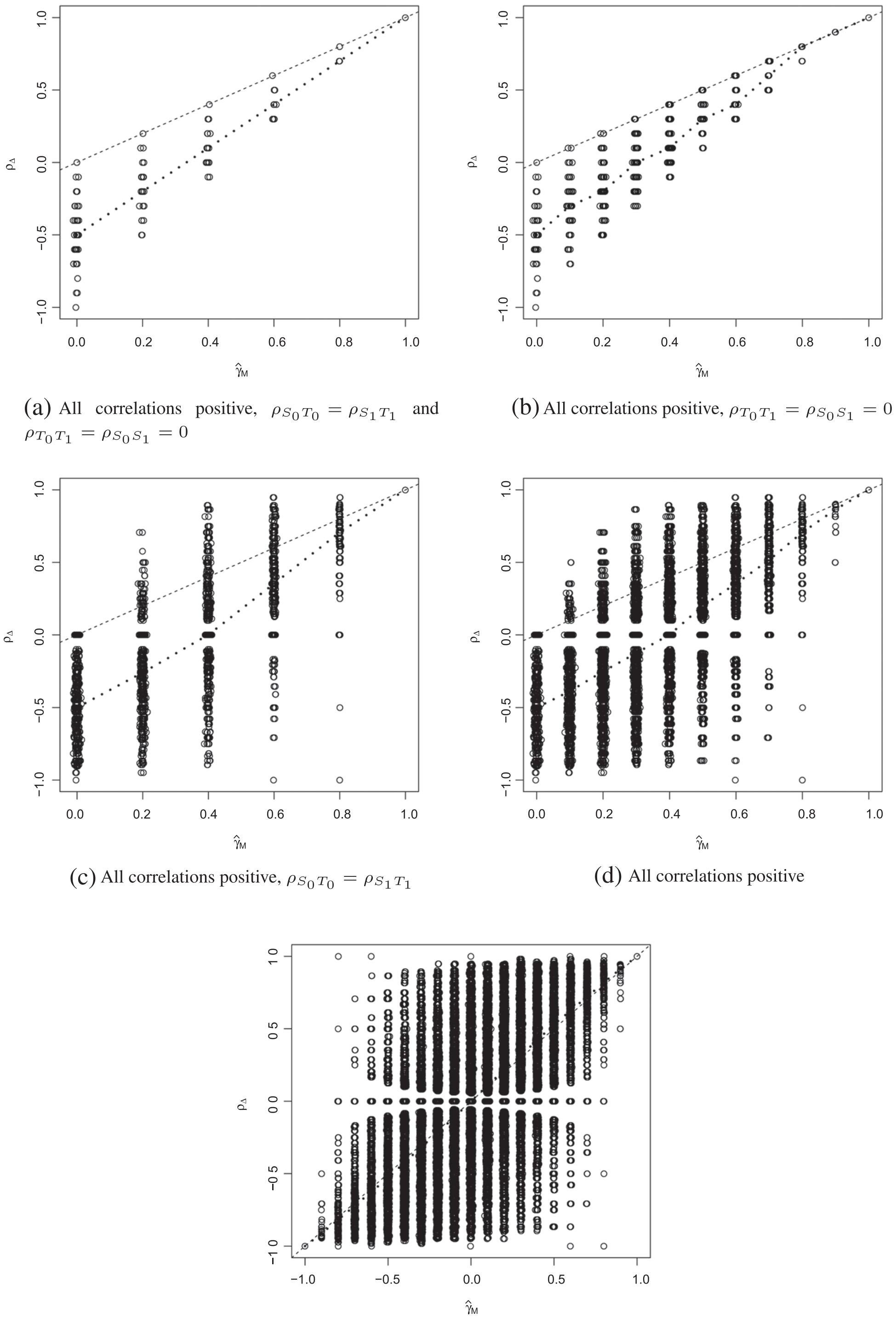
Data were generated based on the theoretical model introduced in Section 2 for $\boldsymbol{Y}_j$ and assuming that $\mu = (0,\ 0,\ 0,\ 0)'$, $\sigma_{T_0T_0} = \sigma_{T_1T_1} = \sigma_T = 1$ and $\sigma_{S_0S_0} = \sigma_{S_1S_1} = \sigma_S = 1$. The number of subjects was fixed at 1000, and for all correlations in (3) the grid of values $G = \{-1, -0.80, \ \dots, \ 1\}$ was considered. This led to a total of $6^{11}\boldsymbol{\Sigma}$ matrices, of which only 173,945 were positive definite ($\boldsymbol{\Sigma}_k$). For each of these positive definite matrices, 50 matrices $\mathbf{C}_{kp}$ ($k = 1, \dots, 173,945$ and $p = 1, \dots, 50$) containing the values of the counterfactuals $T_{1j}$, $T_{0j}$, $S_{1j}$, and $S_{0j}$ for each of the 1000 subjects $j$ were generated. Next, the components of the treatment indicator vector $\mathbf{Z}_{kp}$ were independently sampled from a binomial distribution with success probability 0.50. Finally, using the matrices $\mathbf{C}_{kp}$ and the corresponding vectors $\mathbf{Z}_{kp}$, data sets $\mathbf{F}_{kp}$ were constructed containing the observable variables $T_j$, $S_j$, and $Z_j$ for each subject.

Based on the positive definite matrices $\Sigma_k$, expression (3) was used to compute $\rho_{\Delta k}$. Further, using the information in each data set $\mathbf{F}_{kp}$, $\gamma_k$ was estimated as $\widehat{\gamma}_{k.} = (1/p)\sum_p \widehat{\gamma}_{kp}$ with $\widehat{\gamma}_{kp} = \text{corr}(S_j, T_j \mid Z_j, \mathbf{F}_{kp})$. To simplify the notation in the following, the subindex $k$ will be omitted in $\widehat{\gamma}_{k.}$ and $\rho_{\Delta k}$, and we will refer to them simply as $\widehat{\gamma}_M$ and $\rho_\Delta$, respectively.

### 3.1. Results

The results will be discussed in five main scenarios and considering only the positive definite matrices $\Sigma_k$: (i) all correlations are positive, the correlations between the true and surrogate endpoint are equal in both treatment conditions ($\rho_{S_0T_0} = \rho_{S_1T_1} = \gamma$), and the potential outcomes for both endpoints are uncorrelated ($\rho_{S_0S_1} = \rho_{T_0T_1} = 0$); (ii) all correlations are positive and the potential outcomes for both endpoints are uncorrelated ($\rho_{S_0S_1} = \rho_{T_0T_1} = 0$); (iii) all correlations are positive and the correlations between the true and surrogate endpoint are equal in both treatment conditions ($\rho_{S_0T_0} = \rho_{S_1T_1} = \gamma$); (iv) all correlations are positive; and (v) all correlations vary unrestrictedly.

The results of the simulations in scenarios (i) and (ii) are shown in Figure 1(a)–(b) where the dashed lines indicate perfect agreement ($\rho_\Delta = \widehat{\gamma}_M$) and the dotted lines give the median of $\rho_\Delta$ for every value of $\widehat{\gamma}_M$. The graphs are strikingly similar, hinting on the fact that the assumption of independence



(a) All correlations positive, $\rho_{S_0T_0} = \rho_{S_1T_1}$ and $\rho_{T_0T_1} = \rho_{S_0S_1} = 0$

(b) All correlations positive, $\rho_{T_0T_1} = \rho_{S_0S_1} = 0$

(c) All correlations positive, $\rho_{S_0T_0} = \rho_{S_1T_1}$

(d) All correlations positive

(e) Correlations unrestricted in $G$

**Figure 1.** $\rho_\Delta$ against $\widehat{\gamma}_M$. The dashed lines indicate perfect agreement between $\rho_\Delta$ and $\widehat{\gamma}_M$. The dotted lines show the median of $\rho_\Delta$ as a function of $\widehat{\gamma}_M$.

between the potential outcomes likely is the most important factor determining the results. In addition, both graphs clearly follow the pattern implied by inequality (4). Indeed, in scenario (i), (4) takes the simpler form $|\rho_\Delta - \gamma| \leqslant \sqrt{\left(1-\gamma^2\right)}$, and therefore, the difference between $\widehat{\gamma}_M$ and ICA should be close to 0 when the former is sufficiently large. Interestingly, even though scenario (ii) goes beyond the scope of (4), it still exhibits the same behavior, that is, $\rho_\Delta \approx \widehat{\gamma}_M$ when $\widehat{\gamma}_M$ (with $\widehat{\gamma}_M$ giving now the average of the estimated correlations between the true and surrogate endpoints in both treatment groups) is large. This general trend is confirmed in both scenarios by the monotonic relationship between the median of $\rho_\Delta$ and $\widehat{\gamma}_M$ (dotted line). However, in spite of this general trend, when $\widehat{\gamma}_M$ is moderate or low, the adjusted association only offers a poor approximation of ICA. For example, when $\widehat{\gamma}_M \approx 0.60$, $\rho_\Delta$ approximately ranged between 0.30 and 0.60 in both scenarios, indicating that the relationship between $\widehat{\gamma}_M$ and $\rho_\Delta$ may be largely affected by the unidentifiable parameters.

Nonetheless, in all cases, $\widehat{\gamma}_M$ roughly gives an upper bound for $\rho_\Delta$, and therefore, if all correlations are positive and $\rho_{S_0S_1} = \rho_{T_0T_1} = 0$, then the estimated adjusted association may offer valuable information about ICA. In fact, a low $\widehat{\gamma}_M$ would give evidence of a weak ICA, whereas a large $\widehat{\gamma}_M$ could be reasonably interpreted as a good approximation of $\rho_\Delta$ and, hence, an indication of a large ICA. In general, $\text{corr}(\rho_\Delta, \widehat{\gamma}_M) = 0.885/0.875$ in scenarios (i) and (ii), confirming the usefulness of $\widehat{\gamma}_M$ to get information about $\rho_\Delta$. Summarizing, if the assumptions underlying setting (ii) are considered approximately valid, then the observable correlation between both endpoints could be used to draw tentative conclusions about the unidentifiable association between the individual casual effects.

The results of the simulations in scenario (iii) are shown in Figure 1(c). As compared with scenarios (i) and (ii), the correlation between $\rho_\Delta$ and $\widehat{\gamma}_M$ has decreased substantially with $\text{corr}(\rho_\Delta, \widehat{\gamma}_M) = 0.725$, and the variability of $\rho_\Delta$, for a given value of $\widehat{\gamma}_M$, has increased considerably. Actually, only those values of $\widehat{\gamma}_M$ very close to unity are informative regarding ICA. For example, even when $\widehat{\gamma}_M$ is as large as 0.80, $\rho_\Delta$ approximately ranged between $-1.00$ and 0.95.

Figure 1(d) covers the setting in which all correlations are positive. Although scenario (iv) is more general than the one characterized by inequality (4), the same general pattern is recognizable here. Indeed, large values of $\widehat{\gamma}_M$ seem to be associated with large values of $\rho_\Delta$ (dotted line), but this association is rather weak with $\text{corr}(\rho_\Delta, \widehat{\gamma}_M) = 0.632$. Like in scenario (iii), $\widehat{\gamma}_M$ does not offer an upper bound for $\rho_\Delta$ any longer — even though the inequality $\rho_\Delta \leqslant \widehat{\gamma}_M$ seems to hold for most cases, in particular when $\widehat{\gamma}_M$ is high. Indeed, the percentage of cases for which $\rho_\Delta \leqslant \widehat{\gamma}_M$ goes from 60% when $\widehat{\gamma}_M = 0.60$ to 97% when $\widehat{\gamma}_M = 0$.

A comparison with the findings in scenarios (i)–(ii) and (iii)–(iv) indicates that the correlations between the potential outcomes for the true and surrogate endpoints ($\rho_{S_0S_1}$, $\rho_{T_0T_1}$) have an important impact on the relationship between the adjusted association and ICA. To further examine the impact of these correlations on the proportion of cases where $\rho_\Delta$ no longer provides an upper bound for $\widehat{\gamma}_M$ in scenarios (iii) and (iv), consider Table I. As can be seen, this proportion is small when both the correlations between the potential outcomes are low and increases when at least one correlation between the potential outcome is large. Nonetheless, even in the worst case scenario, $\widehat{\gamma}_M$ provides an upper bound for $\rho_\Delta$ in the majority of cases. For example, when $\rho_{S_0S_1} = 0.80$ and $\rho_{T_0T_1} = 0$ in scenario (iii), $\widehat{\gamma}_M$ is below $\rho_\Delta$ in 75% of the cases.

Finally, Figure 1(e) shows the results in the most general setting (v) when all correlations are allowed to vary unrestrictedly. Interestingly, in this completely general scenario, $\widehat{\gamma}_M$ is always approximately equal to the median of $\rho_\Delta$, indicating that the observable correlation between both endpoints is now rather uninformative regarding ICA. In fact, as one would expect, only when both endpoints are almost deterministically related ($\widehat{\gamma}_M = \pm 1$) the observable correlation brings useful information about $\rho_\Delta$.

**Table I.** Proportion of runs in scenarios (iii) and (iv) in which $\rho_\Delta$ is larger than $\widehat{\gamma}_M$ as a function of $\rho_{S_0S_1}$ and $\rho_{T_0T_1}$.

| | | Scenario (iii) | | | | | Scenario (iv) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | $\rho_{T_0T_1}$ | | | | | $\rho_{T_0T_1}$ | | |
| | | 0 | 0.20 | 0.40 | 0.60 | 0.80 | 0 | 0.20 | 0.40 | 0.60 | 0.80 |
| $\rho_{S_0S_1}$ | 0 | 0.022 | 0.044 | 0.084 | 0.128 | 0.232 | 0.048 | 0.054 | 0.093 | 0.128 | 0.178 |
| | 0.20 | 0.055 | 0.068 | 0.106 | 0.117 | 0.177 | 0.057 | 0.075 | 0.106 | 0.142 | 0.166 |
| | 0.40 | 0.074 | 0.106 | 0.121 | 0.196 | 0.176 | 0.092 | 0.108 | 0.129 | 0.175 | 0.159 |
| | 0.60 | 0.130 | 0.128 | 0.207 | 0.157 | 0.158 | 0.128 | 0.141 | 0.171 | 0.160 | 0.192 |
| | 0.80 | 0.250 | 0.161 | 0.176 | 0.175 | 0.150 | 0.184 | 0.155 | 0.164 | 0.197 | 0.162 |

## 4. The multiple-trial setting

A stated in Section 2, in the STS neither ICA nor ECA are identifiable from the data. Indeed, even though the expected causal effects $(\alpha, \beta)$ are identifiable, the relative effect only provides information about ECA under strong and unverifiable assumptions. The assumption most frequently made by data analysts when working with *RE* is that, over the population of trials, $E(\beta_i|\alpha_i) = RE \times \alpha_i$, that is, the expected causal effects satisfy the regression through the origin equation $\beta_i = RE \times \alpha_i + \varepsilon_i$. However, regression through the origin has often been surrounded by controversy because of the paradoxical results it can produce, like negative coefficients of determination [15].

A way out of this problem is to abandon the STS in favor of the multiple-trial setting in order to estimate ECA. Such meta-analytic methods have been proposed by different authors [4, 16–18]. In the following sections, we will focus on the approach introduced by Buyse *et al.* [4] for normally distributed endpoints.

### 4.1. Expected causal association: a meta-analytic approach

In the meta-analytic framework, it is assumed that data from $i = 1, 2, ..., N$ clinical trials, or other relevant clustering units, are available and $n_i$ patients are enrolled in the $i$th trial. In addition, let $T_{ij}$, $S_{ij}$, and $Z_{ij}$ denote the true endpoint, the surrogate endpoint, and the treatment indicator variable for subject $j$ in randomized trial $i$, respectively. Buyse *et al.* [4] considered the following hierarchical model

$$\begin{cases} S_{ij} = \mu_S + m_{Si} + (\alpha + a_i)Z_{ij} + \varepsilon_{Sij} \\ T_{ij} = \mu_T + m_{Ti} + (\beta + b_i)Z_{ij} + \varepsilon_{Tij} \end{cases}, \tag{5}$$

where $\mu_S$ and $\mu_T$ are the fixed intercepts for $S$ and $T$, $m_{Si}$ and $m_{Ti}$ are the corresponding random intercepts, $\alpha$ and $\beta$ are the fixed treatment effects for $S$ and $T$, and $a_i$ and $b_i$ are the corresponding random effects. The error terms $\varepsilon_{Sij}$ and $\varepsilon_{Tij}$ are assumed to be zero-mean normally distributed with covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_{SS} & \sigma_{ST} \\ \sigma_{ST} & \sigma_{TT} \end{pmatrix}. \tag{6}$$

Further, the vector of the random effects $(m_{Si}, m_{Ti}, a_i, b_i)'$ is assumed to be zero-mean normally distributed with covariance matrix $\mathbf{D}$. Using the previous notation, the expected causal effects of $Z$ on $S$ and $T$ in trial $i$ equal $\alpha_i = \alpha + a_i$ and $\beta_i = \beta + b_i$, respectively. Model (5) implies that $\mu'_{\Delta i} = (\alpha_i,\ \beta_i)$ follows also a normal distribution with mean $\bar{\mu}_\Delta = (\bar{\alpha},\ \bar{\beta})'$ and covariance matrix

$$\mathbf{D}_\Delta = \begin{pmatrix} d_{aa} & d_{ab} \\ d_{ab} & d_{bb} \end{pmatrix}. \tag{7}$$

Matrix (7) conveys all the information regarding the association between the expected causal effects, and therefore, ECA can be defined as

$$R_{\text{trial}} = \text{corr}\left(\alpha_i,\ \beta_i\right) = \frac{d_{ab}}{\sqrt{d_{aa}d_{bb}}}. \tag{8}$$

Expression (8) can be seen as a special case of the general measure used by Buyse *et al.* [4] to quantify the so-called trial-level surrogacy. These authors complemented the trial-level surrogacy with the so-called individual-level surrogacy, which is defined as the treatment-corrected and trial-corrected correlation between $S$ and $T$

$$R_{\text{ind}} = \text{corr}\left(\varepsilon_{Sij},\ \varepsilon_{Tij}\right) = \frac{\sigma_{ST}}{\sqrt{\sigma_{SS}\sigma_{TT}}}.$$

### 4.2. Individual causal association: a meta-analytic approach

Although individual causal effects are mostly used in a STS context, they could be employed in a meta-analytic framework as well. To this end, let us extend model (1) by assuming that $\boldsymbol{\Delta}_{ij} \sim N\left(\boldsymbol{\mu}_{\Delta i}, \boldsymbol{\Sigma}_\Delta\right)$ and $\boldsymbol{\mu}_{\Delta i} \sim N\left(\bar{\boldsymbol{\mu}}_\Delta, \boldsymbol{D}_\Delta\right)$, where $\boldsymbol{\Delta}_{ij}$ is the vector of individual causal effects for patient $j$ in trial $i$, $\boldsymbol{\mu}_{\Delta i}$ is the vector of expected causal effects in trial $i$, and $\boldsymbol{\Sigma}_{\Delta}$ and $\boldsymbol{D}_{\Delta}$ are defined as before. Based on this model, Alonso *et al.* [12] defined the meta-analytic individual causal association (MICA) as $\rho_M = \text{corr}\left(\Delta_{Tij}, \Delta_{Sij}\right)$ and showed that

$$\rho_M = \frac{\sqrt{d_{bb}d_{aa}}R_{\text{trial}} + \sqrt{\left(\sigma_{T_0T_0} + \sigma_{T_1T_1} - 2\sqrt{\sigma_{T_0T_0}\sigma_{T_1T_1}}\rho_{T_0T_1}\right)\left(\sigma_{S_0S_0} + \sigma_{S_1S_1} - 2\sqrt{\sigma_{S_0S_0}\sigma_{S_1S_1}}\rho_{S_0S_1}\right)}\rho_{\Delta}}{\sqrt{\left[d_{bb} + \sigma_{T_0T_0} + \sigma_{T_1T_1} - 2\sqrt{\sigma_{T_0T_0}\sigma_{T_1T_1}}\rho_{T_0T_1}\right]\left[d_{aa} + \sigma_{S_0S_0} + \sigma_{S_1S_1} - 2\sqrt{\sigma_{S_0S_0}\sigma_{S_1S_1}}\rho_{S_0S_1}\right]}}, \tag{9}$$

where $\rho_{\Delta}$ is computed using (2). If homoscedasticity holds, that is, $\sigma_{T_0T_0} = \sigma_{T_1T_1} = \sigma_T$ and $\sigma_{S_0S_0} = \sigma_{S_1S_1} = \sigma_S$, then (9) takes the simpler form

$$\rho_M = \frac{\sqrt{d_{bb}d_{aa}}R_{\text{trial}} + 2\sqrt{\sigma_T\sigma_S\left(1 - \rho_{T_0T_1}\right)\left(1 - \rho_{S_0S_1}\right)}\rho_{\Delta}}{\sqrt{\left(d_{bb} + 2\sigma_T\left(1 - \rho_{T_0T_1}\right)\right)\left(d_{aa} + 2\sigma_S\left(1 - \rho_{S_0S_1}\right)\right)}}, \tag{10}$$

with $\rho_{\Delta}$ as given in (3). Expressions (9)–(10) clearly show the complex interplay between the metrics of surrogacy used in the meta-analytic and single-trial contexts. Unlike $\rho_{\Delta}$, which only assesses the validity of the surrogate in a single and fixed population (internal validity), $\rho_M$ evaluates its validity across similar but different populations (external validity). Actually, Equation (10) shows that $\rho_M$ is a weighted sum of a within-trial contribution given by $\rho_{\Delta}$ and a between-trial contribution given by $R_{\text{trial}}$. Which of these two elements is the dominating factor in (10) depends on the relative size of the between-trial variability, as quantified by $d_{bb}$ and $d_{aa}$, and the within-trial variability, as quantified by $\sigma_T$ and $\sigma_S$. Nonetheless, like $\rho_{\Delta}$, $\rho_M$ also suffers from identifiability issues that can be addressed, as before, using a simulation-based sensitivity analysis.

Using expression (10) and assuming $\rho_{T_0T_1} = \rho_{S_0S_1} = 0$, Alonso *et al.* [12] heuristically showed that choosing a surrogate with a large and positive $R_{\text{trial}}$ and $R_{\text{ind}}$ may likely lead to a surrogate with a large and positive $\rho_M$. As a consequence, these authors concluded that the general strategy followed in the meta-analytic paradigm may largely be compatible with the approach that one would follow in the causal inference framework and that surrogates successfully evaluated using the former may be appealing to proponents of the latter. Nonetheless, the validity of this conclusion when the independence assumption does not hold has still to be confirmed. In the following section, this and other important issues will be explored via simulation.

## 5. The multiple-trial setting: simulation study

In the simulations, the number of trials, or other relevant clustering units, was fixed at $N = 150$, each involving $n_i = 20$ subjects. Although this choice may seem unnatural at first sight, settings with a large number of clustering units and a relatively small number of observations per unit are frequently found in practice. In fact, given that the actual number of available clinical trials is often insufficient to apply the multiple-trial methods, researchers frequently resort to alternative clustering units, such as the hospitals where the patients are treated in.

The trial-specific random effects were sampled as $(m_{Si}, m_{Ti}, a_i, b_i) \sim N(0, \mathbf{D})$ with the variances of the random intercepts equal to 1 and the covariances between the random intercepts and between the random intercepts and the random treatment effects all equal to zero. Two sets of values were used for the variances of the treatment random effects: $d_{aa} = d_{bb} = 25$, indicative of a large between-trial variability, and $d_{aa} = d_{bb} = 0.10$ representing homogeneous trials. The covariance between $(a_i, b_i)$ was calculated as $d_{ab} = R_{\text{trial}}\sqrt{d_{aa}d_{ab}}$ with $R_{\text{trial}} = \{0, 0.30, 0.60, 0.90\}$.

Further, within each trial, data were generated based on the theoretical model introduced in Section 2 for $\boldsymbol{Y}_j$ with $\mu = (0, 0, 0, 0)'$, $\sigma_S = \sigma_T = 0.10$ and $\sigma_S = \sigma_T = 25$. For all correlations in (10), the grid of values $G = \{-0.90, -0.60, -0.30, 0, 0.30, 0.60, 0.90\}$ was considered. This led to a total of $6^7$ matrices of which only the 17033 positive definite ones $(\Sigma_k)$ were retained. Notice that, compared with what was the case in the STS, a less narrow grid of values $G$ was used here for the correlations in (10) to keep the simulation time feasible. Using each of the previous positive definite matrices the vector of pseudo-potential outcomes $(T^*_{0ij}, T^*_{1ij}, S^*_{0ij}, S^*_{1ij})$ was generated for each of the 20 subjects in trial $i$. The potential outcomes for the true endpoint were computed as $T_{0ij} = T^*_{0ij} + m_{Ti}$, $T_{1ij} = T^*_{1ij} + m_{Ti} + \beta + b_i$ (with $\beta = 0$) and for the surrogate as $S_{0ij} = S^*_{0ij} + m_{Si}$, $S_{1ij} = S^*_{1ij} + m_{Si} + \alpha + a_i$ (with $\alpha = 0$). Subsequently, the treatment indicator variable $Z_{ij}$ for each subject was sampled from a binomial distribution with success probability 0.50. Next, using the vectors of potential outcomes and the treatment indicator variables, the data sets for each trial $\mathbf{F}_{ik}$ were constructed containing the observable variables $T_{ij}$, $S_{ij}$, and $Z_{ij}$. For each positive definite covariance matrix $\Sigma_k$, the previous steps were repeated 50 times leading to the final data sets $\mathbf{F}_{ikp}$ (with $i = 1, \dots, 150$, $k = 1, \dots, 17033$, and $p = 1, \dots, 50$).

Finally, using the information in $\mathbf{F}_{ikp}$, the $\widehat{R}_{\text{trial}}$ and $\widehat{R}_{\text{ind}}$ were computed based on the reduced, univariate, fixed-effect approach [19]. For each $i$ and $k$ combination, the means of the estimates $\widehat{R}_{\text{trial}}$ and $\widehat{R}_{\text{ind}}$ were computed over $p$, and they are referred to as $R_{trial_M}$ and $R_{indiv_M}$ in the following section. Expression (10) was used to compute the true value of $\rho_M$ in all settings.

### 5.1. *Results*

The same scenarios that were considered in the STS (Section 3.1) are evaluated here. For succinctness, only the results of scenarios (iv) and (v) are detailed here. The results of scenarios (i)–(iii), which were very similar to those of scenario (iv), are provided in the online Appendix (Part 1).

Tables II and III show the means of the $\rho_M$ true values as a function of $R_{\text{trial}}$ and $\rho_\Delta$ in scenarios (iv) and (v), respectively. Furthermore, together with the mean, $\rho_M$ the mean estimates for $R_{trial_M}$ and $R_{indiv_M}$ are given between brackets $(\bar{R}_{trial_M}, \bar{R}_{indiv_M})$.

Some interesting patterns regarding the relation between $\rho_\Delta$, $R_{\text{trial}}$, and $\rho_M$ emerge from these results. In both scenarios (iv) and (v), the analysis of the relationship between $\rho_M$ and $(R_{\text{trial}}, \rho_\Delta)$ shows that MICA is mainly determined by $R_{\text{trial}}$ when the trial-level variability is substantially larger than the individual-level variability, and $\rho_\Delta$ is the most influential factor when the individual-level variability is substantially larger than the trial-level variability (Tables II and III). This result is in line with the theory, as can be seen from the following re-parametrization of (10)

$$\rho_M = \frac{\sqrt{\lambda_T \lambda_S} R_{\text{trial}} + 2\sqrt{\left(1 - \rho_{T_0T_1}\right)\left(1 - \rho_{S_0S_1}\right)}\rho_\Delta}{\sqrt{\lambda_T \lambda_S + 2\lambda_T\left(1 - \rho_{T_0T_1}\right) + 2\lambda_S\left(1 - \rho_{T_0T_1}\right) + 4\left(1 - \rho_{T_0T_1}\right)\left(1 - \rho_{S_0S_1}\right)}},$$

where $\lambda_T = d_{bb}/\sigma_T$ and $\lambda_S = d_{aa}/\sigma_S$; and, in the limit, $\lim_{\substack{\lambda_T \to \infty \\ \lambda_S \to \infty}} \rho_M = R_{\text{trial}}$ and $\lim_{\substack{\lambda_T \to 0 \\ \lambda_S \to 0}} \rho_M = \rho_\Delta$. Importantly, the previous expression and the results in Tables II and III show that a surrogate successfully evaluated in a single-trial setting (large $\rho_\Delta$) may fail to classify as a good surrogate in a meta-analytic context (low $\rho_M$) when the trial-level heterogeneity is much larger than the individual-level variability (large $\lambda_S = \lambda_T$) and the expected causal association is low (small $R_{\text{trial}}$). For example, as shown in Tables II and III, the mean $\rho_M$ may be close to zero even though $\rho_\Delta$ is larger than 0.75 when $R_{trial_M}$ is close to zero, and the trial-level heterogeneity is much larger than the individual-level variability ($\lambda_S$ and $\lambda_T$ large). It is thus important to emphasize that a single trial evaluation will always be incomplete, unless one is only interested in the validity of the surrogate in the specific population studied in the trial (internal validity).

To examine the impact of the unidentifiable correlations in (10) on $\rho_M$, consider Figure 2(a)–(c). In these figures, the $\rho_M$ true values are shown in the settings where $R_{trial} = 0.90$, $\rho_{S_0T_0} = \rho_{S_1T_1} = 0.60$, and the correlations between the potential outcomes equal 0 or 0.30 for the different trial-level and individual-level variance components. As can be seen in Figure 2(a), $\rho_M$ is close to $R_{trial}$ when the trial-level variability is substantially larger than the individual-level variability — irrespective of the magnitude of the values of the correlations between the potential outcomes (i.e., the variability in $\rho_M$ is small). In contrast, the unidentifiable correlations have a profound impact on $\rho_M$ when the trial-level variability is substantially smaller than the individual-level variability (Figure 2(b)). In particular, higher values of $\rho_{T_0T_1}$ and $\rho_{S_0S_1}$ lead to higher $\rho_M$ values, whereas lower values of $\rho_{T_1S_0}$ and $\rho_{T_0S_1}$ lead to lower $\rho_M$ values. A similar pattern is observed when the trial-level and individual-level variability components are of the same magnitude, although the impact of the unidentifiable correlations on $\rho_M$ becomes smaller in the latter scenario because the relative importance of the trial-level component increases (Figure 2(c)). Similar results were obtained when other values of $R_{trial}$, $d_{aa}$, $d_{bb}$, and unidentifiable correlations were considered (data not shown due to space constraints).

**Table II.** Scenario (iv), all correlations positive. Mean $\rho_M$ as a function of $R_{trial}$ and $\rho_\Delta$ when (a) $d_{aa} = d_{bb} = 25$ and $\sigma_S = \sigma_T = 0.10$, (b) $d_{aa} = d_{bb} = 0.10$ and $\sigma_S = \sigma_T = 25$, and (c) $d_{aa} = d_{bb} = 25$ and $\sigma_S = \sigma_T = 25$. Between brackets $(\bar{R}_{trial_M}, \bar{R}_{indiv_M})$.

| | $R_{trial_M}$ | | | |
|---|---|---|---|---|
| $\rho_\Delta$ | 0 | 0.30 | 0.60 | 0.90 |
| | (a) $d_{aa} = d_{bb} = 25$ and $\sigma_S = \sigma_T = 0.10$ $(\lambda_S = \lambda_T = 250)$ | | | |
| [−1, −0.75] | −0.003 (0.222, 0.220) | 0.296 (0.350, 0.222) | 0.594 (0.615, 0.221) | 0.893 (0.894, 0.221) |
| (−0.75, −0.45] | −0.003 (0.227, 0.203) | 0.296 (0.357, 0.203) | 0.594 (0.608, 0.203) | 0.893 (0.899, 0.203) |
| (−0.45, −0.15] | −0.002 (0.229, 0.250) | 0.297 (0.358, 0.251) | 0.595 (0.605, 0.248) | 0.893 (0.895, 0.250) |
| (−0.15, 0.15] | 0.001 (0.224, 0.347) | 0.299 (0.355, 0.348) | 0.597 (0.605, 0.349) | 0.896 (0.896, 0.349) |
| (0.15, 0.45] | 0.002 (0.226, 0.449) | 0.300 (0.357, 0.449) | 0.598 (0.605, 0.448) | 0.897 (0.896, 0.450) |
| (0.45, 0.75] | 0.003 (0.231, 0.542) | 0.301 (0.353, 0.540) | 0.600 (0.609, 0.541) | 0.898 (0.897, 0.541) |
| (0.75, 1] | 0.004 (0.221, 0.616) | 0.302 (0.360, 0.618) | 0.601 (0.615, 0.617) | 0.899 (0.897, 0.618) |
| | (b) $d_{aa} = d_{bb} = 0.10$ and $\sigma_S = \sigma_T = 25$ $(\lambda_S = \lambda_T = 0.004)$ | | | |
| [−1, −0.75] | −0.818 (0.277, 0.222) | −0.816 (0.271, 0.221) | −0.814 (0.290, 0.219) | −0.813 (0.302, 0.219) |
| (−0.75, −0.45] | −0.558 (0.254, 0.203) | −0.557 (0.266, 0.205) | −0.556 (0.278, 0.203) | −0.555 (0.291, 0.205) |
| (−0.45, −0.15] | −0.279 (0.279, 0.249) | −0.278 (0.285, 0.251) | −0.277 (0.309, 0.251) | −0.276 (0.321, 0.251) |
| (−0.15, 0.15] | 0.006 (0.345, 0.349) | 0.008 (0.373, 0.348) | 0.010 (0.383, 0.348) | 0.011 (0.408, 0.347) |
| (0.15, 0.45] | 0.293 (0.431, 0.450) | 0.294 (0.444, 0.450) | 0.295 (0.467, 0.450) | 0.296 (0.482, 0.450) |
| (0.45, 0.75] | 0.592 (0.502, 0.534) | 0.593 (0.526, 0.533) | 0.595 (0.541, 0.533) | 0.596 (0.570, 0.532) |
| (0.75, 1] | 0.880 (0.656, 0.702) | 0.881 (0.675, 0.699) | 0.882 (0.700, 0.697) | 0.883 (0.730, 0.699) |
| | (c) $d_{aa} = d_{bb} = 25$ and $\sigma_S = \sigma_T = 25$ $(\lambda_S = \lambda_T = 1)$ | | | |
| [−1, −0.75] | −0.369 (0.224, 0.220) | −0.211 (0.364, 0.219) | −0.052 (0.594, 0.217) | 0.106 (0.861, 0.221) |
| (−0.75, −0.45] | −0.285 (0.226, 0.204) | −0.149 (0.347, 0.205) | −0.012 (0.592, 0.202) | 0.125 (0.858, 0.204) |
| (−0.45, −0.15] | −0.158 (0.218, 0.251) | −0.031 (0.357, 0.248) | 0.096 (0.592, 0.249) | 0.223 (0.862, 0.250) |
| (−0.15, 0.15] | 0.004 (0.223, 0.348) | 0.156 (0.359, 0.350) | 0.307 (0.596, 0.349) | 0.459 (0.869, 0.349) |
| (0.15, 0.45] | 0.167 (0.227, 0.447) | 0.294 (0.364, 0.448) | 0.422 (0.592, 0.450) | 0.549 (0.872, 0.449) |
| (0.45, 0.75] | 0.282 (0.226, 0.532) | 0.428 (0.360, 0.534) | 0.573 (0.602, 0.532) | 0.719 (0.877, 0.532) |
| (0.75, 1] | 0.456 (0.231, 0.696) | 0.592 (0.382, 0.697) | 0.728 (0.598, 0.698) | 0.864 (0.886, 0.697) |

**Table III.** Scenario (v), correlations unrestricted in $G$. Mean $\rho_M$ as a function of $R_{trial}$ and $\rho_\Delta$ when (a) $d_{aa} = d_{bb} = 25$ and $\sigma_S = \sigma_T = 0.10$, (b) $d_{aa} = d_{bb} = 0.10$ and $\sigma_S = \sigma_T = 25$, and (c) $d_{aa} = d_{bb} = 25$ and $\sigma_S = \sigma_T = 25$. Between brackets ($\bar{R}_{trial_M}$, $\bar{R}_{indiv_M}$).

| | $R_{trial_M}$ | | | |
|---|---|---|---|---|
| $\rho_\Delta$ | 0 | 0.30 | 0.60 | 0.90 |
| | (a) $d_{aa} = d_{bb} = 25$ and $\sigma_S = \sigma_T = 0.10$ ($\lambda_S = \lambda_T = 250$) | | | |
| [−1, −0.75] | −0.005 (0.226, −0.373) | 0.292 (0.353, −0.373) | 0.590 (0.605, −0.373) | 0.887 (0.896, −0.373) |
| (−0.75, −0.45] | −0.004 (0.225, −0.277) | 0.293 (0.357, −0.276) | 0.591 (0.607, −0.276) | 0.888 (0.896, −0.276) |
| (−0.45, −0.15] | −0.002 (0.226, −0.138) | 0.295 (0.357, −0.138) | 0.593 (0.607, −0.138) | 0.891 (0.896, −0.138) |
| (−0.15, 0.15] | 0.001 (0.225, 0.001) | 0.298 (0.358, 0.002) | 0.595 (0.608, 0.002) | 0.893 (0.896, 0.002) |
| (0.15, 0.45] | 0.002 (0.226, 0.141) | 0.300 (0.356, 0.141) | 0.597 (0.606, 0.141) | 0.895 (0.896, 0.141) |
| (0.45, 0.75] | 0.004 (0.227, 0.280) | 0.302 (0.356, 0.279) | 0.600 (0.607, 0.279) | 0.897 (0.896, 0.279) |
| (0.75, 1] | 0.006 (0.224, 0.383) | 0.304 (0.356, 0.383) | 0.601 (0.608, 0.384) | 0.899 (0.896, 0.384) |
| | (b) $d_{aa} = d_{bb} = 0.10$ and $\sigma_S = \sigma_T = 25$ ($\lambda_S = \lambda_T = 0.004$) | | | |
| [−1, −0.75] | −0.824 (0.427, −0.373) | −0.823 (0.411, −0.373) | −0.822 (0.399, −0.374) | −0.821 (0.388, −0.373) |
| (−0.75, −0.45] | −0.574 (0.341, −0.276) | −0.573 (0.330, −0.277) | −0.572 (0.316, −0.277) | −0.571 (0.308, −0.276) |
| (−0.45, −0.15] | −0.291 (0.257, −0.138) | −0.291 (0.251, −0.139) | −0.290 (0.247, −0.138) | −0.289 (0.241, −0.138) |
| (−0.15, 0.15] | 0.003 (0.232, 0.001) | 0.004 (0.233, 0.001) | 0.005 (0.234, 0.001) | 0.006 (0.239, 0.001) |
| (0.15, 0.45] | 0.299 (0.259, 0.141) | 0.300 (0.263, 0.142) | 0.300 (0.274, 0.141) | 0.301 (0.286, 0.141) |
| (0.45, 0.75] | 0.592 (0.343, 0.273) | 0.593 (0.352, 0.274) | 0.594 (0.368, 0.274) | 0.595 (0.383 0.274) |
| (0.75, 1] | 0.850 (0.448, 0.421) | 0.850 (0.462, 0.419) | 0.851 (0.477, 0.420) | 0.852 (0.493, 0.419) |
| | (c) $d_{aa} = d_{bb} = 25$ and $\sigma_S = \sigma_T = 25$ ($\lambda_S = \lambda_T = 1$) | | | |
| [−1, −0.75] | −0.494 (0.230, −0.373) | −0.380 (0.331, −0.372) | −0.266 (0.566, −0.373) | −0.152 (0.832, −0.372) |
| (−0.75, −0.45] | −0.360 (0.228, −0.277) | −0.253 (0.336, −0.277) | −0.146 (0.566, −2.777) | −0.040 (0.836, −0.276) |
| (−0.45, −0.15] | −0.182 (0.224, −0.138) | −0.075 (0.341, −0.138) | 0.031 (0.573, −0.138) | 0.137 (0.843, −0.138) |
| (−0.15, 0.15] | 0.002 (0.223, 0.002) | 0.114 (0.344, 0.001) | 0.226 (0.580, 0.001) | 0.338 (0.850, 0.002) |
| (0.15, 0.45] | 0.187 (0.226, 0.140) | 0.293 (0.351, 0.141) | 0.399 (0.585, 0.141) | 0.506 (0.857, 0.141) |
| (0.45, 0.75] | 0.361 (0.224, 0.273) | 0.471 (0.355, 0.274) | 0.581 (0.590, 0.273) | 0.691 (0.863, 0.273) |
| (0.75, 1] | 0.538 (0.224, 0.419) | 0.643 (0.360, 0.418) | 0.747 (0.600, 0.419) | 0.851 (0.871, 0.419) |

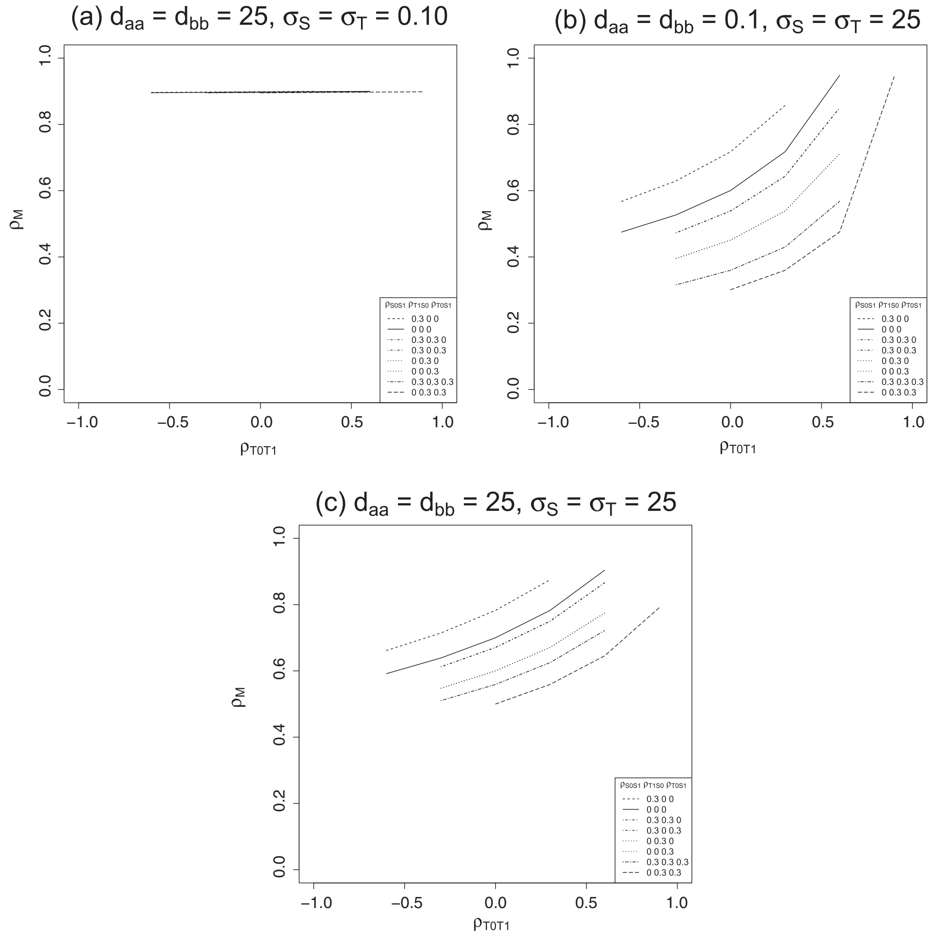


**Figure 2.** $\rho_M$ true values as a function of $\rho_{T_0T_1}$, $\rho_{S_1T_0}$, $\rho_{S_0T_1}$, and $\rho_{S_0S_1}$ in the settings where (a) $d_{aa} = d_{bb} = 25$ and $\sigma_S = \sigma_T = 0.10$, (b) $d_{aa} = d_{bb} = 0.10$ and $\sigma_S = \sigma_T = 25$, and (c) $d_{aa} = d_{bb} = 25$ and $\sigma_S = \sigma_T = 25$. $R_{trial}$ was fixed at 0.90 and $\rho_{T_0S_0} = \rho_{S_1T_1} = 0.60$.

Further, Tables II and III show the relationship between $\rho_M$ and $(\bar{R}_{trial_M}, \bar{R}_{indiv_M})$. As can be seen, in scenario (iv) higher values of $\bar{R}_{trial_M}$ and $\bar{R}_{indiv_M}$ are typically related to large values of $\rho_M$. This indicates that a surrogate that is successfully evaluated at the trial-level and individual-level in the meta-analytic framework will typically also successfully pass the validation exercise based on individual causal effects. The same holds in scenario (v), though $\bar{R}_{indiv_M}$ tends to be low to moderate even when $\rho_\Delta$ is high. For example, even when $\rho_\Delta$ is higher than 0.75, the mean $\bar{R}_{indiv_M}$ is only about 0.40 — irrespective of the magnitude of $R_{trial}$ or the ratio of the trial-level and individual-level variability components. This finding is fully in line with the results detailed in Section 3.1, where it was shown that $\rho_\Delta$ and $\gamma$ are only loosely related in scenario (v). Indeed, as can be seen in Figure 1(e), the $\widehat{\gamma}_M$ values range between −0.799 and 0.999 in the fully general scenario when $\rho_\Delta$ is higher than 0.75 (with a mean of 0.399).

## 6. Case study

In this section, we introduce data from a meta-analysis of five double-blind randomized clinical trials, comparing the effect of risperidone and conventional antipsychotic agents for the treatment of schizophrenia. The data consisted of $N_{total} = 2128$ patients who, depending on the trial, were treated for a period of 4–8 weeks. In psychiatry, several measures can be considered to assess a patient's global condition and in these clinical trials three rating scales were administered to each patient at the start and the end of the study. The Clinical Global Impression (CGI; [20]) is generally accepted by practitioners as a reliable clinical measure of patient status. This is a 7-grade scale used by the treating physician to characterize how well a subject has improved. Another useful and sufficiently sensitive assessment scale is the Positive and Negative Syndrome Scale (PANSS; [21]). 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; [22]) is a subscale of PANSS including only 18 items.

Even though this is not a standard situation for surrogate validation due to the lack of a clear *gold* standard, we consider two primary measures (true endpoints): the Clinical Global Impression scale, which is the one that has the clearest clinical interpretation, and PANSS, which is arguably the most complete and reliable instrument. The main idea was to evaluate if a simpler and, therefore, easier to administer scale like BPRS could be reliably used as a substitute for CGI (scale that requires medical expertise) and/or the more complex PANSS scale. Thus, surrogacy analyses were conducted to examine (i) whether the change in the BPRS score (= BPRS score after the treatment — BPRS score at the start of the treatment) is a good surrogate for the change in the PANSS score, and (ii) whether the change in the BPRS score is a good surrogate for the change in the CGI score. These questions were addressed in both the single-trial and multiple-trial settings. To simplify the exposition in the following sections, the names of the endpoints (BPRS, PANSS and CGI) will be loosely used to refer to the change in score between the beginning and the end of the study for each scale.

As stated before, our meta-analysis contained only five trials. However, at the second level of the hierarchical structure of the data, information was also available regarding the psychiatrists who treated the patients. Hence, we used the $N = 198$ psychiatrists within trial as the basis for our analyses in the multiple-trial setting. This is convenient because the number of trials itself may be too small for a meta-analytic evaluation [1]. The number of patients per cluster (psychiatrists) ranged from $n_i = 1$ to 52 and, when the multiple-trial methods were used, the data of clusters in which only one type of the treatment was administered and/or the surrogate/true endpoint was constant were excluded.

The R package *Surrogate* that accompanies this paper (freely available at CRAN) allows for the evaluation of surrogate endpoints using the meta-analytic and the causal-inference methods previously discussed. For succinctness, in the next sections, only a summary of the main results is given and no reference to the software is made. However, in the Web Appendix (Part II), a more comprehensive analysis is provided and the use of the R package *Surrogate* is explained in detail.

### 6.1. *The single-trial setting*

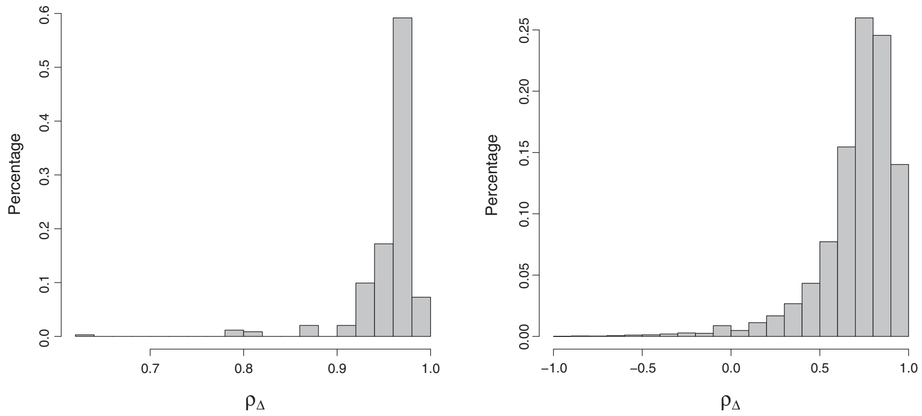
*6.1.1. Change in the BPRS as a surrogate for change in the PANSS.* As expected, the observable association between $S$ = BPRS and $T$ = PANSS was large in the control and experimental groups, equaling $\widehat{\rho}_{S_0T_0} = 0.960$ ($CI_{95\%} = [0.956,\ 0.963]$) and $\widehat{\rho}_{S_1T_1} = 0.964$ ($CI_{95\%} = [0.961,\ 0.967]$), respectively. When the information from both groups was combined, the adjusted association was estimated as $\widehat{\gamma} = 0.963$ ($CI_{95\%} = [0.960,\ 0.966]$).

The results in Section 3.1 indicated that such a strong association between both endpoints ($\widehat{\gamma} = 0.963$) may be indicative for a strong and positive correlation between the corresponding individual causal effects as well. As noted earlier, $\rho_\Delta$ is not identifiable from the data, and therefore, the simulation-based sensitivity analysis introduced in Section 2 was used. To that end, the observable correlations $\rho_{S_0T_0}$ and $\rho_{S_1T_1}$ were fixed at their estimated values and for all the unidentifiable correlations the grid of values $G = \{-1,\ -0.90,\ \ldots,\ 1\}$ was considered. This led to a total of $21^4$ matrices of which only 343 were positive definite. Figure 3(a) depicts the behavior of the estimates of $\rho_\Delta$ across these plausible correlation matrices. In line with the results of the simulations, most of the estimated $\rho_\Delta$ values were large, with 95% of them exceeding 0.901 (*Mean* = 0.957, *SD* = 0.037, range [0.620; 0.994]). These large values suggest that the individual causal effects on the simpler BPRS scale convey a substantial amount of information about the individual causal effects on the more complex PANSS.

The validity of a putative surrogate can also be evaluated from a prediction perspective. Actually, in practice, one would like to predict the individual causal effect of $Z$ on $T$ in patient $j$ ($\Delta_{Tj}$) based on the individual causal effect of $Z$ on $S$ ($\Delta_{Sj}$). Model (1) introduced in Section 2 implies

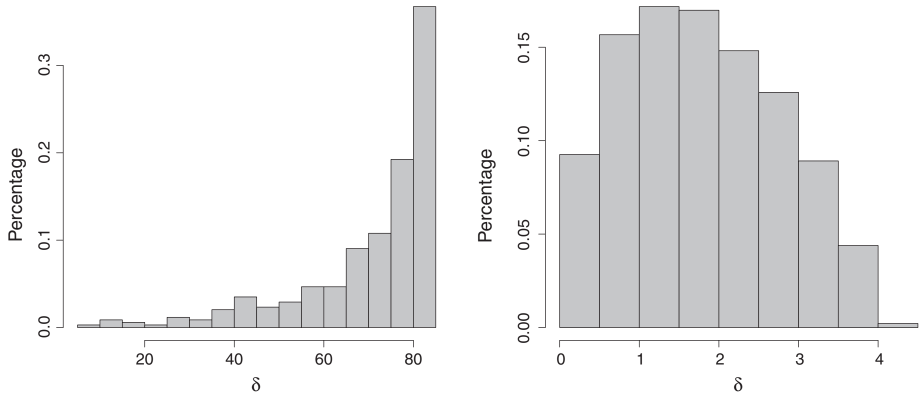
$$\Delta_{Tj}|\Delta_{Sj} \sim N\left[g\left(\Delta_{Sj}\right), \sigma_{\Delta_T}\left(1-\rho_\Delta^2\right)\right],$$

where $g\left(\Delta_{Sj}\right) = \beta + \sqrt{\frac{\sigma_{\Delta_T}}{\sigma_{\Delta_S}}}\rho_\Delta(\Delta_{Sj}-\alpha)$, $\sigma_{\Delta_T} = \sigma_{T_0T_0} + \sigma_{T_1T_1} - 2\sqrt{\sigma_{T_0T_0}\sigma_{T_1T_1}}\rho_{T_0T_1}$ and $\sigma_{\Delta_S} = \sigma_{S_0S_0} + \sigma_{S_1S_1} - 2\sqrt{\sigma_{S_0S_0}\sigma_{S_1S_1}}\rho_{S_0S_1}$ with $\alpha$, $\beta$, the variance components and the correlations as before. The prediction mean squared error (PMSE) can be quantified as



(a) $S$ = BPRS and $T$ = PANSS (b) $S$ = BPRS and $T$ = CGI

**Figure 3.** Analyses of the Risperidone clinical trial data in the single-trial causal-inference framework. Histograms of $\rho_\Delta$ for the analyses where (a) the change in the Brief Psychiatric Rating Scale (BPRS) was the surrogate and the change in the Positive and Negative Syndrome Scale (PANSS) the true endpoint, and (b) the change in the BPRS was the surrogate and the change in the Clinical Global Impression (CGI) the true endpoint.



(a) $S$ = BPRS and $T$ = PANSS (b) $S$ = BPRS and $T$ = CGI

**Figure 4.** Analyses of the Risperidone clinical trial data in the single-trial causal-inference framework. Histograms of $\delta$ for the analyses where (a) the change in the Brief Psychiatric Rating Scale (BPRS) was the surrogate and the change in the Positive and Negative Syndrome Scale (PANSS) the true endpoint, and (b) the change in the BPRS was the surrogate and the change in the Clinical Global Impression (CGI) the true endpoint.

$$\delta = E\left[\left(\Delta_{Tj} - g\left(\Delta_{Sj}\right)\right)^2\right] = \sigma_{\Delta_T}\left(1-\rho_\Delta^2\right).$$

The previous equation illustrates that PMSE is the product of two elements: ICA and the variance of the individual causal effect on the true endpoint. Even though ICA depends on the surrogate, $\sigma_{\Delta_T}$ only depends on the true endpoint and the treatment, and hence, the search for a good surrogate endpoint may not be viable in some situations. In fact, if $\sigma_{\Delta_T}$ is very large, then one would need to find a surrogate with an ICA almost equal to one to be able to make meaningful predictions.

Using the previously obtained estimates from the simulation-based algorithm, one can also construct a frequency distribution for the PMSE (Figure 4(a)). In 95% of the cases the PMSE is 37.118 or higher (with a maximum value of 83.940). Thus, using the individual causal effect on BPRS, the individual causal effect on PANSS can be predicted with a prediction error between about 6 and 9 points (notice that PANSS ranges between −102 and 81 points). All in all, one may conclude that the BPRS seems to be a good surrogate for the PANSS.

*6.1.2. Change in the BPRS as a surrogate for change in the CGI.* The correlations between $S$ = BPRS and the second true endpoint $T$ = CGI were $\widehat{\rho}_{S_0T_0} = 0.734$ ($CI_{95\%}$ = [0.714, 0.753]) and $\widehat{\rho}_{S_1T_1} = 0.739$

($CI_{95\%} = [0.720,\ 0.758]$) in the control and experimental group, respectively. When the information from both groups was pooled the adjusted association $\widehat{\gamma}$ equaled 0.738 ($CI_{95\%} = [0.718,\ 0.757]$).

Notice that the association between the surrogate and true endpoint is now substantially smaller than in the previous scenario. Similarly, the mean value for the estimates of $\rho_{\Delta}$ was much lower, and their variability and range were larger (*Mean* = 0.711, *SD* = 0.221, range [−0.943, 0.999]), as can be seen in Figure 3(b). Clearly, the values assumed for the unidentifiable correlations between the potential outcomes seem to have a much larger impact on the results in this setting. Therefore, one may conclude that the validity of the BPRS as a surrogate for the CGI is not clearly established in this analysis and the results are sensitive to the assumptions regarding the unidentifiable correlations.

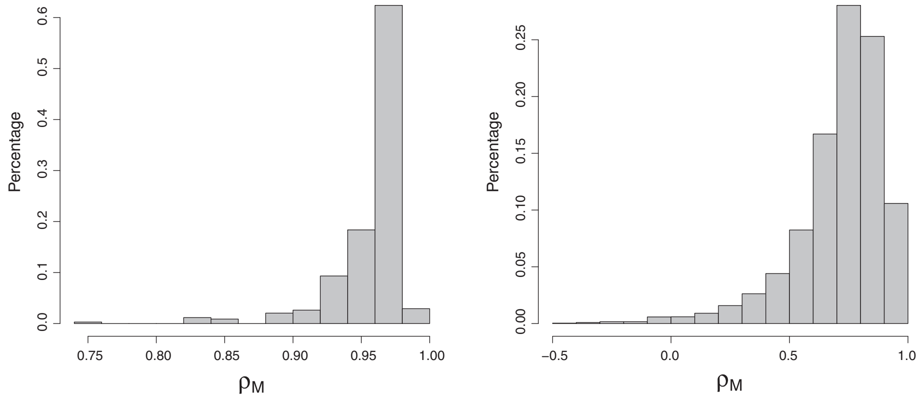
The frequency distribution for the PMSE is shown in Figure 4(b). As can be seen, most of the $\delta$ values lay in the [0.5; 3.5] interval. The individual causal effect on CGI can thus be predicted, using the individual causal effect on BPRS, with a prediction error between about 0.7 and 1.9 points (notice that CGI ranges between 1 and 7 points).

### 6.2. *The multiple-trial setting*

*6.2.1. Change in the BPRS as a surrogate for change in the PANSS* . The results from the meta-analytic approach largely resembled the ones obtained in the single-trial setting. Indeed, the expected causal association was very large with an estimated value of $\widehat{R}_{\text{trial}} = 0.959$ ($CI_{95\%} = [0.944,\ 0.970]$). Similarly, the correlation between both endpoints after adjusting for treatment and treating physician was very large, $\widehat{R}_{\text{ind}} = 0.963$ ($CI_{95\%} = [0.949,\ 0.973]$). Therefore, from a meta-analytic perspective, one may conclude that there is evidence of a very strong association at the two levels of the hierarchy, that is,at the patient and treating physician levels.

As previously stated, one can also use individual causal effects in a meta-analytic context. The validity of the surrogate can then be assessed using the meta-analytic individual causal association. To that effect, a simulation-based sensitivity analysis was again implemented, and the grid of values $G = \{-1,\ -0.90,\ ...,\ 1\}$ was used for all the unidentified correlations in (9). Figure 5(a) illustrates that most of the estimated $\rho_M$ values were high, with 95% of them larger than 0.909 (*Mean* = 0.958). Furthermore, the variability and range of the distribution of $\rho_M$ was smaller than what was the case in the single-trial setting (*SD* = 0.029, range [0.752, 0.985]) — as expected, because trial-level information is taken into account in the computation of $\rho_M$. The small variability in $\rho_M$ indicates that the results are not sensitive to the assumptions regarding the unidentified correlations. Thus, the present analysis seems to indicate that the BPRS may be considered an appropriate surrogate for the PANSS.

*6.2.2. Change in the BPRS as a surrogate for change in the CGI.* When analyzing the validity of $S$ = BPRS as a surrogate for $T$ = CGI, more moderate results were found. Indeed, the expected causal



(a) $S$ = BPRS and $T$ = PANSS

(b) $S$ = BPRS and $T$ = CGI

**Figure 5.** Analyses of the Risperidone clinical trial data in the multiple-trial causal-inference framework. Histograms of $\rho_M$ for the analyses where (a) the change in the Brief Psychiatric Rating Scale (BPRS) was the surrogate and the change in the Positive and Negative Syndrome Scale (PANSS) the true endpoint, and (b) the change in the BPRS was the surrogate and the change in the Clinical Global Impression (CGI) the true endpoint.

**Table IV.** Analyses of the Risperidone clinical trial data. Summary statistics for $\rho_\Delta$ and $\rho_M$ in the settings where the sampling variability in the estimation of $\rho_{T_0S_0}$ and $\rho_{T_1S_1}$ is not accounted for (left) and is accounted for (right).

| | | | Sampling variability not accounted for | | | Sampling variability accounted for | | |
|---|---|---|---|---|---|---|---|---|
| | | | | $\rho_\Delta$ | | | $\rho_\Delta$ | |
| | *S* | *T* | Mean | *SD* | Range | Mean | *SD* | Range |
| ICA | BPRS | PANSS | 0.9567 | 0.0366 | [0.6200; 0.9935] | 0.9566 | 0.0367 | [0.6122; 0.9940] |
| | BPRS | CGI | 0.7108 | 0.2214 | [−0.9429; 0.9996] | 0.7094 | 0.2209 | [−0.9681; 0.9998] |
| | | | | $\rho_M$ | | | $\rho_M$ | |
| | *S* | *T* | Mean | *SD* | Range | Mean | *SD* | Range |
| MICA | BPRS | PANSS | 0.9577 | 0.0285 | [0.7515; 0.9852] | 0.9575 | 0.0286 | [0.7554; 0.9863] |
| | BPRS | CGI | 0.7145 | 0.1892 | [−0.4708; 0.9850] | 0.7133 | 0.1894 | [−0.4845; 0.9877] |

association was estimated as $\widehat{R}_{\text{trial}} = 0.718$ ($CI_{95\%} = [0.630,\ 0.788]$) and the individual level surrogacy as $\widehat{R}_{\text{ind}} = 0.736$ ($CI_{95\%} = [0.652,\ 0.802]$).

As can be learned from Figure 5(b), the estimated values of MICA were much lower and more variable in this scenario with a $Mean = 0.713$, $SD = 0.190$, and a range $[-0.479,\ 0.984]$. Thus, in line with the results in Section 5.1, it can be concluded that the assumptions that one is willing to make regarding the unidentifiable correlations have a strong impact on the conclusion regarding the appropriateness of BPRS as a surrogate for CGI.

### *6.3. Accounting for the sampling variability in the estimation of $\rho_{S_0T_0}$ and $\rho_{S_1T_1}$*

In the analyses detailed earlier, the sampling variability in the estimation of $\rho_{S_0T_0}$ and $\rho_{S_1T_1}$ was not accounted for, that is, these correlations were fixed at their estimated values. To take the imprecision in the estimation of $\rho_{S_0T_0}$ and $\rho_{S_1T_1}$ into account in the sensitivity analysis, these correlations can be sampled from a uniform distribution with (min, max) values equal to the upper and lower bounds of their corresponding 95% CIs. Thus, for each of the $21^4$ matrices that are considered in the analyses, different $\rho_{S_0T_0}$ and $\rho_{S_1T_1}$ values are sampled from these uniform distributions instead of keeping them fixed at their estimated values.

Table IV shows the results. As expected, the range and/or the *SD* of $\rho_\Delta$ and $\rho_M$ tended to be larger when the imprecision in the estimation of $\rho_{S_0T_0}$ and $\rho_{S_1T_1}$ was taken into account, although the effect was very small and did not affect the substantive conclusions of the analyses.

## 7. Discussion

Previous research showed that the meta-analytic and causal-inference frameworks for the evaluation of surrogate endpoints are closely related, but the complexity of this relationship hindered the use of analytic techniques. The first aim of the present study was to further examine the relationship between these frameworks in the single-trial and multiple-trial settings, using simulations and the analysis of a case study.

The results found in the single-trial setting indicate that the relationship between the observable correlation $\gamma$ and ICA, given in (4), is valid even beyond the assumptions used for its derivation. The simulations also seem to confirm the consented view that, in general, a correlate (high $\rho_{S_0T_0}$ and $\rho_{S_1T_1}$) does not make a surrogate. Further, the results help to clarify the assumptions under which the adjusted association may convey substantial information about ICA. Although these assumptions are essentially unverifiable from the data, subject-specific knowledge may help to evaluate their plausibility in a concrete situation.

The results of the simulations are also useful to get a deeper understanding of the complex relationship between MICA ($\rho_M$), ECA ($R_{\text{trial}}$), ICA ($\rho_\Delta$), and the individual level surrogacy ($R_{\text{ind}}$). They seem to extend the heuristic findings introduced in Alonso *et al.* [12], suggesting that a surrogate successfully evaluated in a meta-analytic framework may likely pass a similar validation exercise based on individual causal effects when it can be reasonably assumed that all unidentifiable correlations are positive. In addition, they also confirmed that a validation exercise carried out in a single-trial setting may not be sufficient to establish the validity of the surrogate across similar but different trials.

Overall, the results of the simulations showed that the extent to which the causal-inference and meta-analytic surrogate evaluation paradigms lead to similar conclusions regarding the appropriateness of the surrogate at hand is the result of a complex interplay between different factors, in particular (i) the setting that is used to assess surrogacy (i.e., the single-trial versus multiple-trial setting), (ii) the parameter estimates that are obtained for the identifiable quantities in $\rho_\Delta$ and $\rho_M$ (e.g., the trial-level and individual-level variance components), (iii) the assumptions that are made regarding the correlations between the potential outcomes, and (iv) the assumptions that are made regarding the individual-level variance components (i.e., is homoscedasticity assumed or not). For example, the simulations showed that $\rho_M$ is close to $R_{\text{trial}}$ when the trial-level variability is 'substantially larger' than the individual-level variability. This result is insightful, but it does not allow for specifying a threshold for the ratio of the trial-level and individual-level variance components that guarantees good agreement *in a particular dataset* at hand. The R package *Surrogate* provides a convenient tool in this context, as it allows for the simultaneous examination of the combined impact of these different factors that affect agreement in a straightforward way.

## Acknowledgements

Financial support from the IAP research network #$P7/06$ of the Belgian Government (Belgian Science Policy) is gratefully acknowledged. The research leading to these results has also received funding from the European Seventh Framework programme [FP7 2007–2013] under grant agreement no. 602552.

## References

1. Burzykowski T, Molenberghs G, Buyse M. *The Evaluation of Surrogate Endpoints*. Springer-Verlag: New York, 2005.
2. Baker SG, Kramer BS. Evaluating surrogate endpoints, prognostic markers, and predictive markers: some simple themes. *Clinical Trials* 2015; **12**:299–308.
3. Buyse M, Molenberghs G. The validation of surrogate endpoints in randomized experiments. *Biometrics* 1998; **54**: 1014–1029.
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. Freedman LS, Graubard BI, Schatzkin A. Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine* 1992; **11**:167–178.
6. Molenberghs G, Burzykowski T, Alonso A, Assam P, Tilahun A, Buyse M. A unified framework for the evaluation of surrogate endpoints in mental-health clinical trials. *Statistical Methods in Medical Research* 2010; **19**:205–236.
7. Prentice RL. Surrogate endpoints in clinical trials: definitions and operational criteria. *Statistics in Medicine* 1989; **8**: 431–440.
8. Baker SG, Kramer BS. The risky reliance on small surrogate end point studies when planning a large prevention trial. *Journal of the Royal Statistical Society* 2013; **176**:603–608.
9. Li Y, Taylor JMG, Elliott MR, Sargent D. Causal assessment of surrogacy in a meta-analysis of colorectal cancer trials. *Biostatistics* 2011; **12**:478–492.
10. Conlon AS, Taylor JMG, Elliott MR. Surrogacy assessment using principal stratification when surrogate and outcome measures are multivariate normal. *Biostatistics* 2014; **15**:266–283.
11. Joffe MM, Greene T. Related causal frameworks for surrogate outcomes. *Biometrics* 2009; **65**:530–538.
12. 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.
13. Rubin DB. Estimating causal effects on treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 1974; **66**:688–701.
14. Holland PW. Statistics and causal inference. *Journal of the American Statistical Association* 1986; **81**:945–960.
15. Kutner MH, Nachtsheim CJ, Neter J, Li W. *Applied Linear Statistical Models* (5th ed.) McGraw-Hill: New York, 2005.
16. Albert JM, Ionnidis JP, Reicheldefer P, Conway B, Coombs RW, Crane L, Demasi R, Dixon DO, Flandre P, Hughes MD, Kalish LA, Larntz K, Lin D, Marschner IC, Muñoz A, Murray J, Neaton J, Pettinelli C, Rida W, Taylor JM, Welles SL. Statistical issues for hiv surrogate endpoints: point and counterpoint. *Statistics in Medicine* 1998; **17**:1435–2462.
17. Daniels MJ, Hughes MD. Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine* 1997; **16**:1515–1527.
18. Gail MH, Pfeiffer R, van Houwelingen HC, Carroll RJ. On meta-analytic assessment of surrogate outcomes. *Biostatistics* 2000; **1**:231–246.
19. Tibaldi FS, Cortiñas Abrahantes J, Molenberghs G, Renard D, Burzykowski T, Buyse M, Parmar M, Stijnen T, Wolfinger R. Simplified hierarchical linear models for the evaluation of surrogate endpoints. *Journal of Statistical Computation and Simulation* 2003; **73**:643–658.
20. Guy W. *ECDEU Assessment Manual for Psychopharmacology - Revised*. U.S. Department of Health, Education, and Welfare: Rockville, MD, 1976.
21. 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.
22. Overall J, Gorham D. The brief psychiatric rating scale. *Psychological Reports* 1962; **10**:799–812.

## Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.