Received: 28 March 2020 | Revised: 6 April 2021 | Accepted: 9 May 2021

DOI: 10.1002/pst.2141

MAIN PAPER

WILEY

The individual-level surrogate threshold effect in a causal-inference setting with normally distributed endpoints

Wim Van der Elst[1] | **Ariel Alonso Abad**[2] | **Hans Coppenolle**[1] | **Paul Meyvisch**[3] | **Geert Molenberghs**[2,4]

[1]The Janssen Pharmaceutical companies of Johnson & Johnson, Beerse, Belgium

[2]I-BioStat, KU Leuven, Leuven, Belgium

[3]Galapagos NV, Mechelen, Belgium

[4]UHasselt, Hasselt, Belgium

Correspondence
Wim Van der Elst, The Janssen Pharmaceutical companies of Johnson & Johnson, Belgium.
Email: wim.vanderelst@gmail.com

Abstract

In the meta-analytic surrogate evaluation framework, the trial-level coefficient of determination $\left(R^2_{\text{trial}}\right)$ quantifies the strength of the association between the expected causal treatment effects on the surrogate (S) and the true (T) endpoints. Burzykowski and Buyse supplemented this metric of surrogacy with the surrogate threshold effect (STE), which is defined as the minimum value of the causal treatment effect on S for which the predicted causal treatment effect on T exceeds zero. The STE supplements R^2_{trial} with a more direct clinically interpretable metric of surrogacy. Alonso et al. proposed to evaluate surrogacy based on the strength of the association between the individual (rather than expected) causal treatment effects on S and T. In the current paper, the individual-level surrogate threshold effect (ISTE) is introduced in the setting where S and T are normally distributed variables. ISTE is defined as the minimum value of the individual causal treatment effect on S for which the lower limit of the prediction interval around the individual causal treatment effect on T exceeds zero. The newly proposed methodology is applied in a case study, and it is illustrated that ISTE has an appealing clinical interpretation. The R package surrogate implements the methodology and a web appendix (supporting information) that details how the analyses can be conducted in practice is provided.

KEYWORDS

causal inference, information theory, surrogate threshold effect

1 | INTRODUCTION

The duration, complexity and cost of a clinical trial are substantially affected by the endpoints that are used to assess treatment efficacy.[1–4] The most credible indicator of the response to the new treatment (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 'replacement endpoint' that can be measured earlier, occurs more frequently, is more ethically acceptable, and/or is cheaper. If such a replacement endpoint allows for the accurate prediction of the treatment effect on the true endpoint, it is called a surrogate endpoint.[5–8]

The statistical evaluation of a candidate surrogate is not a trivial endeavour, and different strategies have been developed for this purpose. In the meta-analytic framework, Buyse and Molenberghs[6] proposed to quantify surrogacy based on two metrics. The trial-level coefficient of determination $\left(R^2_{\text{trial}}\right)$ quantifies the strength of the association between the expected causal treatment effects on the true (T) and the surrogate (S) endpoints across different clinical trials. The individual-level coefficient of determination $\left(R^2_{\text{ind}}\right)$ quantifies the strength of the association between S and T across different patients after adjustment for trial- and treatment-effects. Burzykowski and Buyse[3] supplemented these metrics of surrogacy with the surrogate threshold effect (STE), which is defined as the minimum value of the treatment effect on S for which the predicted treatment effect on T is significantly different from zero. The STE is an appealing metric of surrogacy that is complementary to R^2_{trial}, in the sense that it allows for a more direct clinical evaluation of the appropriateness of a candidate surrogate endpoint.

Several authors have argued that understanding the association between *individual* (rather than expected) causal treatment effects on S and T is critical to assess surrogacy.[1,9,10] To this end, Alonso et al.[1] introduced the individual causal association (ICA), which is defined as the correlation between the individual causal treatment effects on S and T (in the setting where both S and T are normally distributed endpoints). In the current paper, an STE-like metric of surrogacy is introduced in the single-trial causal-inference framework. The so-called individual-level STE (ISTE) will be defined as the minimum value of the individual causal treatment effect on S for which the lower limit of the prediction interval around the individual causal treatment effect on T exceeds zero. Similarly as what is the case with the STE, the ISTE allows for a more direct assessment of the clinical usefulness of the candidate S that supplements ICA. Notice that ISTE and the STE differ in terms of their data requirements, that is, one clinical trial is sufficient to compute ISTE whereas multiple clinical trials are needed to compute STE.

The remainder of this paper is organised as follows. In section 2, Rubin's causal-inference model[11] is introduced. In sections 3 and 4, the ISTE is defined and related concepts are explored. In section 5, the identifiability issues that are encountered in estimating ISTE are discussed. In section 6, the newly developed methodology is exemplified in a case study. In section 7, a simulation study is conducted. Finally, some critical comments regarding the newly proposed methodology are given in section 8. The methodology is implemented into the R package surrogate (available for download at CRAN), and a web appendix S1 that accompanies this paper shows how the package can be used to conduct the analyses in practice.

2 | CAUSAL-INFERENCE MODEL AND THE INDIVIDUAL CAUSAL ASSOCIATION

It will be assumed throughout this paper that data were collected on the surrogate (S) and the true endpoint (T) for N patients in a single clinical trial where two treatments are evaluated in a parallel study design. No sub-index for patients will be used to simplify the notation.

Rubin's model for causal inference[11] assumes that each patient has two potential outcomes for T: an outcome T_0 that would be observed under the control treatment ($Z = 0$), and an outcome T_1 that would be observed under the experimental treatment ($Z = 1$). T_0 and T_1 are potential outcomes in the sense that they represent the outcomes of the patient had he or she received the control or the experimental treatment, respectively. Similarly, it is assumed that each patient has two potential outcomes for S, that is, S_0 and S_1. The four-dimensional vector of potential outcomes is then defined as $\boldsymbol{Y} = (T_0, T_1, S_0, S_1)'$ and the corresponding vector of individual causal treatment effects $\boldsymbol{\Delta} = (\Delta T, \Delta S)'$, where $\Delta T = T_1 - T_0$ and $\Delta S = S_1 - S_0$.

The *expected* (or average) causal treatment effects on S and T in the population of interest can be estimated as $E(\boldsymbol{\Delta}) = (\beta, \alpha)$, where $\beta = E(\Delta T)$ and $\alpha = E(\Delta S)$. Rosenbaum and Rubin[12] provided three identifiability conditions under which it is possible to obtain consistent estimators of the expected causal treatment effects. If Y denotes the response of interest and Y_z the potential outcome associated with $Z = z$ then the three identifiability conditions are: (1) Consistency: If $Z = z$ for a given subject then $Y_z = Y$ for that subject, (2) Conditional exchangeability: There is no unmeasured confounding given the data on baseline covariates L, that is, $Y_z \perp Z \,|\, L = l$ for each possible value z of Z and l of L and (3) Positivity: If $f_L(l) \neq 0$ then $f_{Z\,|\,L}(z|\, l) > 0$.[13] It can easily be shown that in randomised clinical trials all conditions hold and the expected causal treatment effects can be estimated as $\beta = E(T|\, Z = 1) - E(T|\, Z = 0)$ and $\alpha = E(S|\, Z = 1) - E(S|\, Z = 0)$, where the conditional expectations are estimated using the observed means in the control and treated groups. The distribution of the vector of potential outcomes $\boldsymbol{Y}$ plays an important role in the surrogate evaluation context. In the following, it will be further assumed that $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = \left(\mu_{T_0}, \mu_{T_1}, \mu_{S_0}, \mu_{S_1}\right)'$ and:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{T_0T_0} & \sigma_{T_0T_1} & \sigma_{T_0S_0} & \sigma_{T_0S_1} \\ \sigma_{T_0T_1} & \sigma_{T_1T_1} & \sigma_{T_1S_0} & \sigma_{T_1S_1} \\ \sigma_{T_0S_0} & \sigma_{T_1S_0} & \sigma_{S_0S_0} & \sigma_{S_0S_1} \\ \sigma_{T_0S_1} & \sigma_{T_1S_1} & \sigma_{S_0S_1} & \sigma_{S_1S_1} \end{pmatrix}.$$

In spite of being a powerful theoretical tool, the concept of potential outcomes raises some important methodological challenges. For example, even though the joint distribution of the subvector (T_z, S_z) is identifiable, these marginals do not fully determine the joint distribution of $\boldsymbol{Y}$ and thus the multivariate normality assumption $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is not verifiable. Notwithstanding this issue, several authors have already used the multivariate normal distribution to model potential outcomes.[1,10,14] Further, it has been established in information theory that the normal distribution has the maximum entropy among all distributions with a specified mean and covariance, and consequently the assumption of normality imposes minimal prior structural constraints beyond these moments.[15] So unless there is strong evidence against the use of the normal model (based on e.g., major violations of bivariate normality for (T_z, S_z) in the dataset at hand), the normality assumption seems to be a sensible choice as it is the least restrictive from all the possible unverifiable distributional assumptions one can chose from.

Given the aforementioned distributional assumptions, the following holds for the vector of individual causal treatment effects:

$$\boldsymbol{\Delta} = \mathbf{A}\boldsymbol{Y} = \begin{pmatrix} T_1 - T_0 \\ S_1 - S_0 \end{pmatrix} \sim N(\boldsymbol{\mu}_\Delta, \boldsymbol{\Sigma}_\Delta), \text{where } \mathbf{A} = \begin{pmatrix} -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix},$$

$\boldsymbol{\Sigma}_\Delta = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^{'}, \boldsymbol{\mu}_\Delta = (\beta, \alpha)^{'}$ with $\beta = E(\Delta T) = \mu_{T_1} - \mu_{T_0}$ and $\alpha = E(\Delta S) = \mu_{S_1} - \mu_{S_0}$. It has been argued that if S is a good surrogate for T, then ΔS should convey a substantial amount of information about ΔT.[1] The amount of uncertainty in ΔT that is expected to be removed when the value of ΔS becomes known is referred to as the mutual information. In the normal setting the concepts of mutual information and correlation are equivalent. In line with these developments, Alonso et al.[1] proposed the use of the ICA to assess surrogacy:

$$\rho_\Delta = \text{corr}(\Delta T, \Delta S) = \frac{\sqrt{\sigma_{T_0T_0}\sigma_{S_0S_0}}\rho_{T_0S_0} + \sqrt{\sigma_{T_1T_1}\sigma_{S_1S_1}}\rho_{T_1S_1} - \sqrt{\sigma_{T_1T_1}\sigma_{S_0S_0}}\rho_{T_1S_0} - \sqrt{\sigma_{T_0T_0}\sigma_{S_1S_1}}\rho_{T_0S_1}}{\sqrt{\left(\sigma_{T_0T_0} + \sigma_{T_1T_1} - 2\sqrt{\sigma_{T_0T_0}\sigma_{T_1T_1}}\rho_{T_0T_1}\right)\left(\sigma_{S_0S_0} + \sigma_{S_1S_1} - 2\sqrt{\sigma_{S_0S_0}\sigma_{S_1S_1}}\rho_{S_0S_1}\right)}}, \quad (1)$$

where ρ_{XY} denotes the correlation between the potential outcomes X and Y.

3 | THE INDIVIDUAL-LEVEL SURROGATE THRESHOLD EFFECT

Based on the causal-inference model detailed in section 2, let us now consider the relation between the expected value of ΔT and ΔS:

$$E(\Delta T|\Delta S) = \gamma_0 + \gamma_1 \Delta S, \quad (2)$$

with $\gamma_0 = \beta - \gamma_1 \alpha$ and $\gamma_1 = \frac{\sigma_{\Delta S \Delta T}}{\sigma_{\Delta S \Delta S}}$, where $\sigma_{\Delta S \Delta T} = \sigma_{T_1S_1} + \sigma_{T_1S_0} + \sigma_{T_0S_1} + \sigma_{T_0S_0}$ and $\sigma_{\Delta S \Delta S} = \sigma_{S_1S_1} + \sigma_{S_0S_0} - 2\sigma_{S_1S_0}$. The upper and lower bounds of the $(1-\zeta)\%$ prediction interval (PI) around $E(\Delta T|\,\Delta S)$ for ΔS_0 equal:

$$l(\Delta S_0) = \gamma_0 + \gamma_1 \Delta S_0 - t_{(1-\zeta/2, N-1)}\sqrt{\text{MSE}\left(1 + \frac{1}{N} + \frac{(\Delta S_0 - \alpha)^2}{\sigma_{\Delta S \Delta S}(N-1)}\right)}, \quad (3)$$

$$u(\Delta S_0) = \gamma_0 + \gamma_1 \Delta S_0 + t_{(1-\zeta/2, N-1)}\sqrt{\text{MSE}\left(1 + \frac{1}{N} + \frac{(\Delta S_0 - \alpha)^2}{\sigma_{\Delta S \Delta S}(N-1)}\right)}, \quad (4)$$

with ΔS_0 is the individual causal treatment effect for a new observation, N is the number of patients in the clinical trial, $t_{(1-\zeta/2,N-1)}$ is the $(1-\zeta)-$th percentile of a t-distribution with $N-2$ degrees of freedom, and MSE(mean squared error) $= \frac{\sum(\Delta T - E(\Delta T|\Delta S_0))}{N-2} = \left(\frac{N-1}{N-2}\right)\left(\sigma_{\Delta T\Delta T} - \frac{\sigma^2_{\Delta T\Delta S}}{\sigma_{\Delta S\Delta S}}\right)$ with $\sigma_{\Delta T\Delta T} = \sigma_{T_1T_1} + \sigma_{T_0T_0} - 2\sigma_{T_1T_0}$. The $(1-\zeta)$% PI corresponds to the interval in which a future observation is expected to fall with the stated level of probability. In the frequentist framework (which is used in the present paper), this probability relates to (hypothetical) repetitions of the same experiment or study. Alternatively, a PI could also be formulated in the Bayesian framework, where it would refer to the interval that contains a future observation with subjective conditional probability $(1-\zeta)$% (for details, see Reference 16).

The value ΔS_0 in Equation (3) for which it holds that $l(\Delta S_0) = 0$ is defined as the individual-level surrogate threshold effect (ISTE). The ISTE can be obtained by setting $l(\Delta S_0)$ equal to 0 in Equation (3) and solving the equation for ΔS_0:

$$\text{ISTE} = \frac{-\sqrt{A - BC} + D}{E}, \tag{5}$$

where,

$$A = \left(-2N\gamma_0\gamma_1\sigma_{\Delta S\Delta S} + 2\gamma_0\gamma_1\sigma_{\Delta S\Delta S} - 2\alpha\text{MSE}t^2_{(1-\zeta/2,N-1)}\right)^2,$$

$$B = 4\left(-N\gamma_1^2\sigma_{\Delta S\Delta S} + \gamma_1^2\sigma_{\Delta S\Delta S} + \text{MSE}t^2_{(1-\zeta/2,N-1)}\right),$$

$$C = -N\gamma_0^2\sigma_{\Delta S\Delta S} + \gamma_0^2\sigma_{\Delta S\Delta S} + \alpha^2\text{MSE}t^2_{(1-\zeta/2,N-1)} +$$

$$\text{MSE}Nt^2_{(1-\zeta/2,N-1)}\sigma_{\Delta S\Delta S} - \frac{\text{MSE}t^2_{(1-\zeta/2,N-1)}\sigma_{\Delta S\Delta S}}{N},$$

$$D = 2N\gamma_0\gamma_1\sigma_{\Delta S\Delta S} - 2\gamma_0\gamma_1\sigma_{\Delta S\Delta S} + 2\alpha\text{MSE}t^2_{(1-\zeta/2,N-1)},$$

and

$$E = 2\left(-N\gamma_1^2\sigma_{\Delta S\Delta S} + \gamma_1^2\sigma_{\Delta S\Delta S} + \text{MSE}t^2_{(1-\zeta/2,N-1)}\right).$$

Figure 1A graphically illustrates how ISTE is determined, that is, ISTE corresponds to the value ΔS_0 for which the lower bound of the PI $l(\Delta S_0)$ equals 0 (black arrow in the figure). When ISTE is large, ΔS_0 should be large in order to conclude a non-zero individual causal treatment effect on T. In such a case, the candidate surrogate may not be useful in practice even though the individual causal treatment effects on S and T are highly correlated (i.e., high ICA). ISTE can thus supplement ICA with important information regarding the (clinical) usefulness of the surrogate.

Notice that for some true endpoints a higher value is indicative for a poorer outcome (e.g., when T is the intensity of pain or level of depressive symptoms), and thus a negative ΔT is indicative for a beneficial treatment effect. In such a situation, ISTE corresponds to the value ΔS_0 for which it holds that $u(\Delta S_0) = 0$, that is, ISTE $= \frac{\sqrt{A-BC}+D}{E}$.

4 | RELATED CONCEPTS

4.1 | Average causal necessity and sufficiency

Based on the principal stratification approach proposed by Frangakis and Rubin,[17] average causal necessity and sufficiency were defined in the following way[18]:

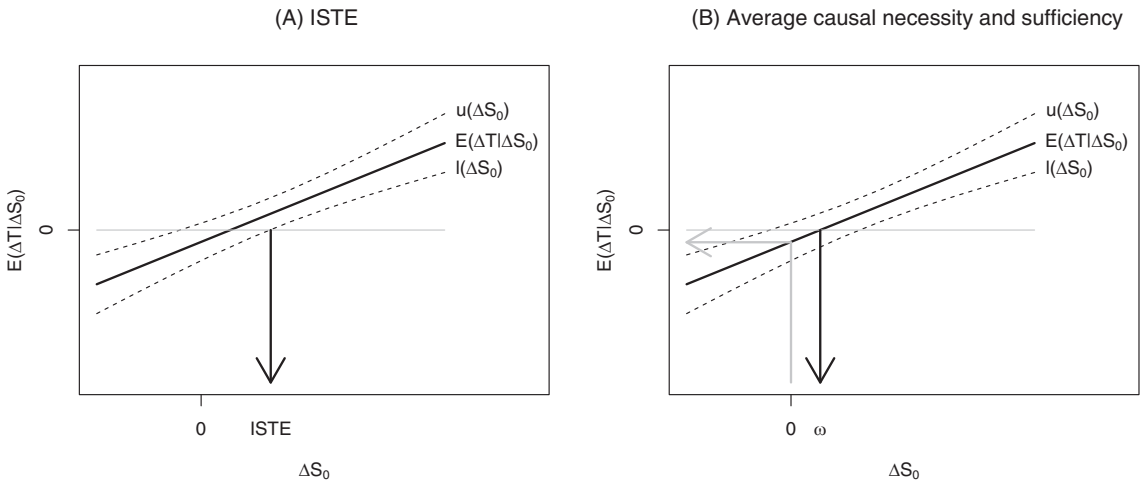

FIGURE 1 Expected ΔT as a function of ΔS_0 (black solid line) and $(1-\zeta)$% PIs (black dashed lines). Panel (A) shows how the ISTE is determined, that is, ISTE is the value ΔS_0 for which it holds that the lower bound of the prediction interval $l(\Delta S_0)$ equals 0 (black arrow). (B) shows w, which is the value ΔS_0 for which it holds that $E(\Delta T|\,\Delta S_0) = 0$ (black arrow). Notice that average causal necessity does not hold in this example because $E(\Delta T|\,\Delta S_0 = 0) \neq 0$ (grey arrow)

$$\text{Average causal necessity}: E(\Delta T|\Delta S_0 = 0) = 0,$$

$$\text{Average causal sufficiency}: \text{ There exists a constant } w \text{ such that } E(\Delta T|\Delta S_0 > w) > 0.$$

Average causal necessity states that in groups of patients who have no individual causal treatment effect on S, the expected causal treatment effect on T should be zero. If average causal necessity holds, γ_0 in Equation (2) is zero.[14] The average causal necessity definition is appealing but may also be restrictive. Indeed, it was shown that even when ICA = 1 (i.e., ΔS and ΔT are deterministically related), average causal necessity may not hold unless further assumptions are made (for details, see Reference 1).

The average causal sufficiency definition states that there is a minimum individual causal treatment effect w on S that guarantees a positive expected causal treatment effect on T. Alonso et al.[1] showed that w exists when ICA is positive. In the above notation and assuming that ICA is positive, w corresponds to $-\frac{\gamma_0}{\gamma_1}$ (see Equation 2). Importantly, even when ICA is positive, there may be individual patients for whom the treatment has no significant impact or even has a negative effect on T.[1] Indeed, the treatment can be expected to be harmful for patients who have an individual causal treatment effect on S for which it holds that $u(\Delta S_0) < 0$ (see Equation 4). Notice that w is always closer to 0 than ISTE because the latter metric considers the prediction error around the expected value of ΔT. This is illustrated in Figure 1B (black arrow). Observe also that average causal necessity does not hold in the example shown in Figure 1, because $E(\Delta T|\,\Delta S_0 = 0) \neq 0$ (see grey arrow).

4.2 | The surrogate predicted function

In the setting where both S and T are binary endpoints, Alonso et al.[19] proposed the surrogate predictive function (SPF). The SPF essentially allows for the determination of the most likely outcome of ΔT for a given value of $\Delta S = \{-1,0,1\}$. A similar idea can be used in the current setting, that is, the expected ΔT value can be computed for a given value of ΔS_0 to get a better understanding of the relation between both (see section 6).

4.3 | The relative effect

ISTE is also connected to the relative effect (RE) that was proposed by Buyse and Molenberghs.[5] The RE is essentially the slope of a regression line between the *expected* (trial-level) causal treatment effects on T and S, that is, $\text{RE} = \frac{\beta}{\alpha}$.

Similarly, the γ_1 parameter in Equation (2) reflects the change in $E(\Delta T|\, \Delta S_0)$ when ΔS_0 increases by one unit. When average causal sufficiency holds (i.e, when γ_0 in Equation (2) equals 0), the RE corresponds to γ_1.

5 | ESTIMATING ISTE

The so-called fundamental problem of causal-inference states that only one of the potential outcomes associated with T and S are observed in practice. As the correlations $\rho_{T_0T_1}$, $\rho_{T_0S_1}$, $\rho_{T_1S_0}$ and $\rho_{S_0S_1}$ in $\mathbf{\Sigma}$ cannot be estimated, the ISTE is not identifiable either. To deal with these identifiability issues, a simulation-based sensitivity analysis can be conducted in which the ISTE is estimated across a set of plausible values for the unidentifiable parameters in $\mathbf{\Sigma}$. Van der Elst et al.[10] proposed an algorithm in which one first specifies a distribution for the unidentified correlations, say $G \sim \text{unif}(-1, 1)$. The use of the latter distribution can be justified based on Laplace's principle of indifference, which essentially states that k possible outcomes should be assigned equal probabilities when no other information is available.[15] That is, if there are no data available suggesting that one outcome is more likely than another one, then each possibility should be assigned a probability equal to $1/k$. In a Bayesian framework, this would correspond to a non-informative prior. Alternatively, the use of the uniform distribution $G \sim \text{unif}(-1, 1)$ can also be justified based on information-theoretic principles, as the uniform distribution on the interval (a, b) is the maximum entropy distribution among all continuous distributions which are supported in the interval (a, b).

In the next step of the algorithm, several covariance matrices $\mathbf{\Sigma}$ are generated by sampling the unidentifiable correlations randomly and independently from G. Under the identifiability conditions described above, the identifiable parameters can be estimated using the data from the control and experimental treatment groups. In the subsequent steps the identifiable parameters are fixed at their estimated values. From the previous matrices only those that are positive definite are retained to calculate the ISTE. The so-obtained vector of values quantifies the ISTE across many plausible 'realities,' that is, across many different scenarios where the assumptions that are made for the unidentified correlations are compatible with the observed data. The general behaviour of ISTE can subsequently be examined, for example, by quantifying the variability and the range of its estimates. In this way, the sensitivity of the results with respect to the unverifiable assumptions (uncertainty) can be evaluated.

In some situations, biological or substantive knowledge may impose reasonable restrictions on the grid G that is considered for the unidentifiable correlations in the algorithm. For example, it might be biologically-sound to assume that some of the unidentified correlations are positive (e.g., see References 10 and 14). In such cases, Laplace's principle of indifference would no longer apply and the grid $G \sim \text{unif}(-1, 1)$ would be replaced by $G \sim \text{unif}(0, 1)$ to appropriately reflect the biological or scientific knowledge when carrying out the sensitivity analysis. The simulation-based approach allows for a straightforward incorporation of such knowledge when it is available. Importantly, any such assumptions are not verifiable based on the data because they relate to unidentified parameters, and thus caution is needed when making these assumptions (see also the results of the simulation study that are described in section 7).

5.1 | Accounting for estimation error

In the approach detailed above, all identifiable quantities are fixed to their estimated values and thus the sampling variability in the estimated parameters is not accounted for. To account for the sampling variability, one can specify distributions for the identifiable parameters—instead of fixing them at their estimated values. For example, uniform distributions can be specified for the identifiable parameters with [min, max] values equal to their estimated upper and lower 95% confidence interval bounds,[20] or alternatively bootstrapped distributions can be used.[15,21] Then random values are drawn from the specified distributions in each run of the algorithm (rather than fixing them to the point estimate).

6 | CASE STUDY: THE RISPERDAL STUDY

6.1 | The dataset

This dataset combines the data that were collected in five double-blind randomised clinical trials. In these trials, the objective was to examine the efficacy of risperidone to treat schizophrenia. Schizophrenia is a mental disease that is hallmarked by hallucinations and delusions.[22]

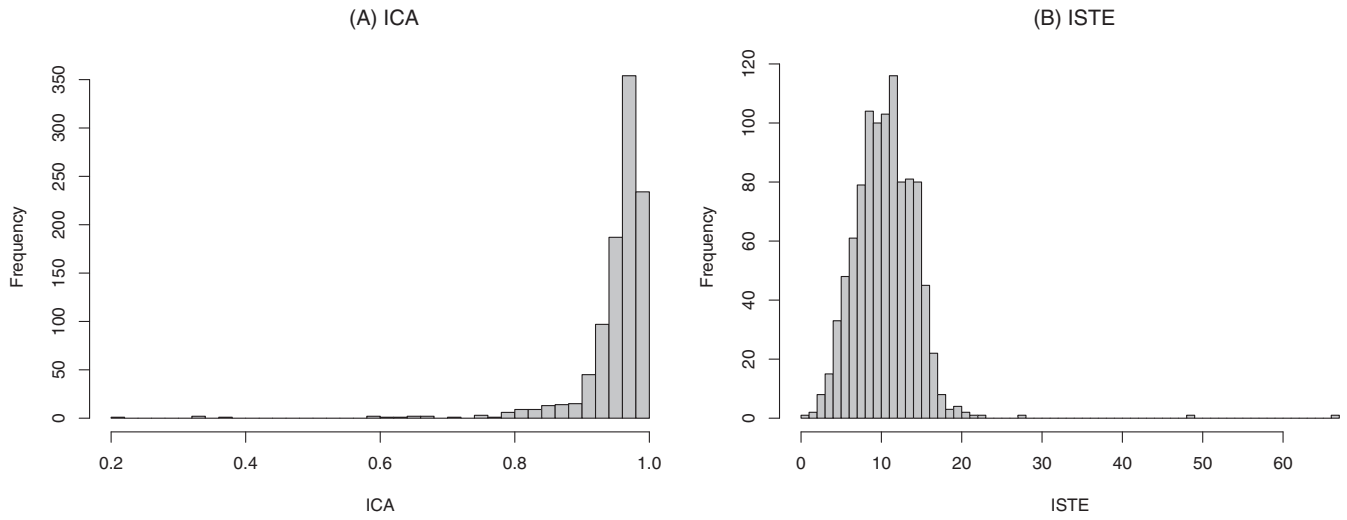

FIGURE 2 Risperdal study. Histogram of (A) ICA (left panel) and (B) ISTE (right panel) across all realities compatible with the observed data

In each trial, the brief psychiatric rating scale (BPRS)[23] and positive and negative syndrome scale (PANSS)[24] were administered. These instruments are clinical rating scales that are routinely used to assess symptom severity in patients with schizophrenia.[25] The patients in the different trials were administered the experimental treatment risperidone or an active control treatment (e.g., haloperidol, levomepromazine, or perphenazine) for 4 to 8 weeks. The main endpoints of interest were S is the change in the BPRS score (= BPRS score at the end of the treatment − BPRS score at the start of the treatment), and T = the change in the PANSS score. A total of 2128 patients participated in the five trials, of whom 1591 patients received risperidone and 537 patients were given an active control.

6.2 | Analysis

To compute ICA and ISTE, a sensitivity analysis is conducted using the grids $G = \{-1, -0.999, ..., 1\}$ for all unidentified correlations. To account for estimation error, the identifiable parameters were sampled from uniform distributions with [min, max] bounds that equalled the 95% confidence intervals around their estimated values. These 95% CIs corresponded to [0.952,0.966] for $\rho_{T_0S_0}$, [0.961,0.968] for $\rho_{T_1S_1}$, [484.462,616.082] for $\sigma_{T_0T_0}$, [514.279,591.062] for $\sigma_{T_1T_1}$, [160.811,204.501] for $\sigma_{S_0S_0}$, [168.989,194.219] for $\sigma_{S_1S_1}$, $[-13.455, -9.489]$ for $E(T_0)$, $[-17.170, -14.860]$ for $E(T_1)$, $[-7.789, -5.503]$ for $E(S_0)$, and $[-9.600, -8.276]$ for $E(S_1)$. In the computation of ISTE, 95% PIs were used.

6.3 | ICA

Figure 2A (left panel) shows a histogram of ICA. As can be seen, ICA is high with mean = 0.9500, median = 0.9655, and 95% of the ICA values exceeding 0.8568. Thus ΔT and ΔS are strongly positively associated, which suggests that S is a good surrogate for T. However, ICA has no direct clinical interpretation and it is thus useful to supplement this metric of surrogacy with the ISTE. For example, ICA does not reflect how large the individual causal treatment effect on S should be to conclude a non-zero individual causal treatment effect on T. In practical terms, one would hope to get values of the individual causal treatment effect on S that can realistically be achieved, given the range of treatment effects on S that are considered to be feasible by medical experts in the field. If the individual causal treatment effects on S that are needed to conclude non-zero individual causal treatment effects on T are too high, the practical usefulness of the candidate surrogate would be low even when ICA is very high.

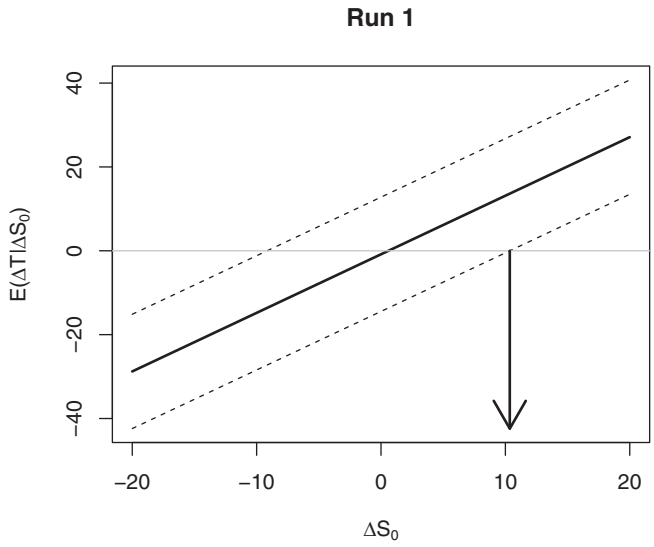

FIGURE 3 Risperdal study. Expected ΔT as a function of ΔS_0 (black solid line) and 95% prediction intervals (black dashed lines) in run 1 of the algorithm. The ISTE in run 1 of the algorithm equals 10.3534 (black arrow)

6.4 | ISTE

The analysis is conducted using 1000 runs of the algorithm described in section 5. Figure 3 illustrates the procedure graphically for the first run of the algorithm, where the black solid line shows $E(\Delta T|\, \Delta S_0)$, the black dashed lines are the 95% PI around the prediction, and the black arrow identifies the ISTE. Here, ISTE = 10.3534. Thus in the first 'reality' that is compatible with the observed data (i.e., the results obtained in the first run of the algorithm), patients who have $\Delta S_0 \geq 10.3534$ are expected to have a significant positive individual causal treatment effect on T. Figure 2 further shows that the treatment is expected to be harmful (i.e., negative ΔT) for patients who have $\Delta S_0 \leq -9.1598$, and that the treatment is not expected to have a significant impact on ΔT (positive or negative) for patients who have $-9.1598 < \Delta S_0 < 10.3534$.

Figure 2B shows the histogram of the ISTE values that are obtained in all 1000 runs of the algorithm. The ISTE values range between 0.1480 and 62.4699, with mean ISTE = 10.6019 and median = 10.6893. This implies that (on average) an individual causal treatment effect of about 10.5 units on S is expected to result in a significant positive individual causal treatment effect on T. Further, 95% of the ISTE values are below 15.7135 and thus an individual causal treatment effect on S that equals about 16 units (or less) is associated with a significant positive individual causal treatment effect on T in most scenarios that are compatible with the observed data. Whether or not it is realistic to have ΔS values of this magnitude for the treatment at hand can be discussed with medical experts in the field. Note that the use of a uniform distribution $G \sim \text{unif}(-1, 1)$ for the unidentified correlations in the sensitivity analysis is similar to assuming a uniform prior in a Bayesian framework, and thus Figure 2B can also be considered an approximation of the posterior ISTE distribution.

6.5 | Average causal necessity and sufficiency

Average causal necessity holds when $E(\Delta T|\, \Delta S_0 = 0) = 0$. Figure 4A shows the histogram of the $E(\Delta T|\, \Delta S_0 = 0)$ values that are obtained in the 1000 runs of the algorithm. The mean $E(\Delta T|\, \Delta S_0 = 0)$ equalled −0.8427, and the figure further shows that $E(\Delta T|\, \Delta S = 0)$ is relatively close to zero in most cases. The average causal necessity condition thus holds approximately in most realities that are compatible with the observed data.

It was furthermore shown that the constant w in $E(\Delta T|\, \Delta S_0 > w) > 0$ exists when ICA is positive[1] (as is the case here). Figure 3B shows the histogram of the w values that are obtained in the different runs of the algorithm. As can be seen, the individual causal treatment effect on T can be expected to exceed 0 for patients who have an individual causal treatment effect on S that ranges between about −2 and 4. Observe that the values of w (Figure 4B) are closer to zero compared to the values of ISTE (see Figure 2B). This is expected because w does not account for prediction error.

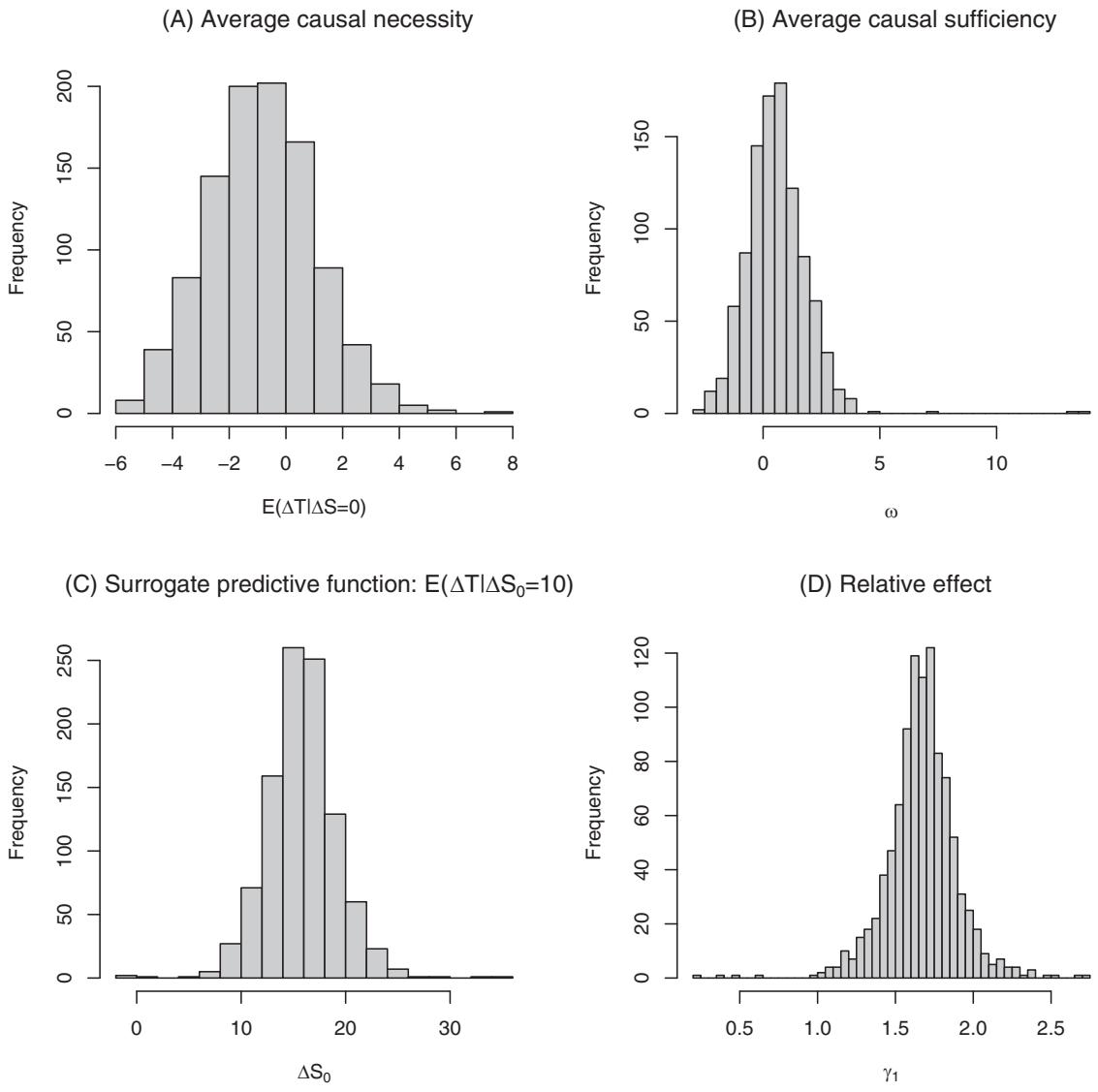

FIGURE 4 Risperdal study. Histograms of (A) $E(\Delta T|\, \Delta S_0 = 0)$ (upper left panel), (B) w for which $E(\Delta T|\, \Delta S_0 > w) = 0$ (upper right panel), (C) $E(\Delta T \,|\, \Delta S_0 = 10$ (bottom left panel), and (D) γ_1 (bottom right panel)

6.6 | The surrogate predictive function

To get a better understanding of the relation between the individual causal treatment effects on S and T, it is insightful to compute the expected value of ΔT for a given ΔS_0 (or a set of values). By means of illustration, consider Figure 4C, which shows the expected ΔT for $\Delta S_0 = 10$. As can be seen, a patient who has an individual causal treatment effect on S that equals 10 is expected to have an individual causal treatment effect on T that ranges between about 5 and 25 units.

6.7 | The relative effect

The estimated expected (trial-level) causal treatment effects on S and T equal $\hat{\alpha} = -1.1461$ and $\hat{\beta} = -2.2716$, respectively. The RE thus equals 1.9820, with 95% confidence interval (approximated using the Delta method) = [1.6610,2.3029]. Figure 4D shows the histogram of the obtained γ_1 values in the sensitivity analysis. The γ_1 parameter reflects the expected change in $E(\Delta T|\, \Delta S_0)$ when ΔS_0 increases by one unit, and it thus has a similar interpretation as RE. If average causal necessity holds (which is approximately the case here, see above), RE corresponds to γ_1. The mean $\gamma_1 = 1.6705$ here, which is indeed close to the RE.

6.8 | Setting biologically plausible constraints

In the above analyses, the grid $G = \{-1, -0.999, ..., 1\}$ was used for all the unidentified correlations in the sensitivity analysis. As noted in section 5, restrictions on G can be imposed when biological or substantive knowledge is available.[10,14] For example, suppose that experts in the field consider it biologically plausible that all identified correlations are positive, that $\rho_{S_0T_1} < \rho_{S_1T_1}$, and that $\rho_{S_0T_1} < \rho_{S_0S_1}$. When the sensitivity analysis is repeated under these constraints (using 1000 runs of the algorithm), the range of ISTE narrows down from [0.1480, 62.4699] (when no assumptions are made) to [0.1480, 37.6450] (when biologically plausible assumptions are made). As expected, the range (i.e., uncertainty) of ISTE is thus reduced when biological information is incorporated in the sensitivity analysis.

Web appendix S1 in Part I of the web appendix, it is exemplified how the case study analyses can be conducted in the surrogate R package. In the current analysis the sampling variability in the estimated parameters was accounted for by sampling from uniform distributions. In the web appendix S1, an alternative bootstrap-based approach to account for sampling variability was used as well. The results of both approaches were essentially the same. Further, a partial check of the multivariate normality assumption was conducted by evaluating the bivariate normality of (S_z, T_z) and its univariate marginals. The latter analysis indicated that there were no major violations of normality.

7 | SIMULATION STUDY

In practice, it is desired to have a surrogate with an ISTE that is as low as possible. ISTE is a complex function of several identifiable and unidentifiable parameters. A simulation study was conducted (1) to identify the conditions under which a low ISTE is obtained, (2) to explore the relation between ICA and ISTE, and (3) to evaluate the coverage of the $l = [\text{minISTE}, \text{maxISTE}]$ interval obtained from the sensitivity analysis (i.e., the percentage of cases in which the true ISTE was included in l).

7.1 | Simulation scenarios

Data were generated based on the theoretical model introduced in section 2, that is, assuming $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. For $\boldsymbol{\mu}$, it was assumed that $\mu_{T_0} = \{0,1\}$, $\mu_{T_1} = \{0,1\}$, $\mu_{S_0} = \{0,1\}$, and $\mu_{S_1} = \{0,1\}$. For $\boldsymbol{\Sigma}$, it was assumed that $\sigma_{S0S0} = \sigma_{S1S1} = \{1, 2\}$, $\sigma_{T0T0} = \sigma_{T1T1} = \{1, 2\}$, $\rho_{T0S0} = \rho_{T1S1} = \{0.5,0.7,0.9\}$, and the grid $G \sim \text{unif}(-1, 1)$ was used for all unidentified correlations. Only scenario's for which the expected causal treatment effects are non-negative are considered here (i.e., scenario's in which it holds that $\alpha = \{0, 1\}$ and $\beta = \{0, 1\}$). In each of the 108 simulation scenarios, a total of 100 positive definite $\boldsymbol{\Sigma}$ were identified. For each of these, a matrix $\boldsymbol{C}_k$ that contains the counterfactuals T_0, T_1, S_0 and S_1 for a total of 1000 patients was generated based on draws from a multivariate normal. Next, the vectors with the treatment indicators $\boldsymbol{Z}_k$ were independently sampled from a binomial distribution with success probability 50%. Finally, based on $\boldsymbol{C}_k$ and $\boldsymbol{Z}_k$, a total of 10,800 datasets $\boldsymbol{F}_k$ were constructed that contained the observable variables S, T and Z for each patient. Based on these datasets, the sensitivity analysis detailed in section 5 was conducted to estimate ISTE and ICA using 100 runs of the algorithm. The identifiable parameters were sampled from uniform distributions with [min, max] values equal to their estimated upper and lower 95% confidence interval bounds. The median ISTE and ICA of the 100 runs for each of the 10,800 generated datasets were retained as the main outcome metrics (that will be referred to as ISTE_M and ICA_M, respectively). The true ISTE was computed based on $\boldsymbol{C}_k$ to evaluate coverage.

7.2 | Results

7.2.1 | Factors that impact ISTE_M

Table 1 shows the mean (SD) of the ISTE_M values in the different simulation scenarios. The results show that a low ISTE_M is mainly obtained in settings (1) where the candidate surrogate is highly correlated with T, (2) where the expected causal treatment effect on S (i.e., α) is smaller than the expected causal treatment effect on T (i.e., β), and (3) where S and T have a small variance in both the treatment conditions.

7.2.2 | Relation between $ISTE_M$ and ICA_M

Figure 5 shows the ICA_M against the $ISTE_M$ in the 10,800 generated datasets across the different simulation scenarios. It can be readily observed that there is a strong negative correlation between $ISTE_M$ and ICA_M, that is, corr(ICA_M, ISTE_M) $= -0.765$. Thus, a high (or low) ICA_M is typically associated with a low (or high) $ISTE_M$. This result is in line with expectations, as the PI around $E(\Delta T|\, \Delta S)$ will be more narrow if ICA_M is higher. Importantly, the simulation also shows that there is substantial variability in the $ISTE_M$ values for a given fixed value of ICA_M. For example, when $ICA_M \approx 0.90$, the $ISTE_M$ values range between about 0 and 3. High $ISTE_M$ values are typically obtained in scenarios where the expected causal treatment effect on S is substantially larger than the expected causal treatment effect on T (i.e., $\alpha = 1$ and $\beta = 0$). Similarly, higher (or lower) ρ_{T0S0} and ρ_{T1S1} are also indicative for a lower (or higher) $ISTE_M$. These findings corroborate the earlier claim that a high ICA (or a high identifiable correlation between S and T) does not necessarily implies that the candidate surrogate will be useful in practice, as the individual causal treatment effect on ΔS that is needed to conclude a significant individual causal treatment effect on ΔT may not be practically feasible.

7.2.3 | Coverage

Table 2 shows the percentage of cases in which the true ISTE was included in the $l = [\min, \max]$ ISTE interval obtained in the sensitivity analysis. As can be seen, coverage is at least 94% across the different simulation scenarios. The overall (averaged) coverage is 97.9%.

7.2.4 | Accounting for uncertainty in the identifiable quantities

In line with Reference 20 the sampling variability in the estimated parameters was accounted for by sampling from uniform distributions with $[\min, \max]$ values equal to the estimated upper and lower 95% confidence interval bounds. Alternative approaches are viable as well, for example, sampling the identifiable parameters from their bootstrapped distributions.[15,21] In the web appendix S1, the simulation study is repeated using this bootstrap-based approach to account for sampling variability. All results were very similar to the ones that were presented here (for details, see Part II of the web appendix S1).

7.2.5 | Impact of making correct and wrong assumptions for the unidentifiable correlations

The unidentifiable correlations were sampled from $G \sim \text{unif}(-1, 1)$, but in some situations biological knowledge may be available. For example, if it is biologically plausible that an unidentified correlation is positive, then the grid $G \sim \text{unif}(0, 1)$ can be used in the sensitivity analysis to reflect this knowledge (see section 5). To further examine the impact of making such assumptions on coverage, the simulation study was extended by considering two additional settings in which either correct or wrong assumptions are made with respect to the unidentifiable correlations. In the first setting, $G \sim \text{unif}(0, 1)$ or $G \sim \text{unif}(-1, 0)$ are used in the sensitivity analysis for true unidentified correlation(s) in $\boldsymbol{\Sigma}$ that are positive or negative, respectively (so correct assumptions are used in the sensitivity analysis). In the second setting, $G \sim \text{unif}(0, 1)$ and $G \sim \text{unif}(-1, 0)$ are used in the sensitivity analysis for true unidentified correlation(s) in $\boldsymbol{\Sigma}$ that are negative or positive, respectively (so wrong assumptions are used). The results showed that coverage was substantially impacted by the validity of the assumptions that were made for the unidentifiable correlations in the sensitivity analysis, that is, overall coverage $= 83.3\%$ versus 98.9% when wrong and correct assumptions are made, respectively. These results thus indicate that the incorporation of biological knowledge can ameliorate the identifiability issues and improve coverage—but this is only the case when the assumptions that are made are correct. Of course, in real life it is not possible to evaluate the validity of these assumptions empirically based on the data at hand (as they relate to unidentifiable parameters), and thus a careful reflection on the plausibility of any such assumptions used in the sensitivity analysis is needed. Further details with respect to these additional simulations can be found in the web appendix S1 Part II.

TABLE 1 Simulation study

Identifiable correlations	Expected causal treatment effects	$\sigma_{T0T0} = \sigma_{T1T1} = 1$ $\sigma_{S0S0} = \sigma_{S1S1} = 1$	$\sigma_{T0T0} = \sigma_{T1T1} = 1$ $\sigma_{S0S0} = \sigma_{S1S1} = 2$	$\sigma_{T0T0} = \sigma_{T1T1} = 2$ $\sigma_{S0S0} = \sigma_{S1S1} = 1$	$\sigma_{T0T0} = \sigma_{T1T1} = 2$ $\sigma_{S0S0} = \sigma_{S1S1} = 2$
$\rho_{T0S0} = \rho_{T1S1} = 0.5$	$\alpha = 0, \beta = 0$	4.0347 (0.4064)	5.7200 (0.6266)	4.0585 (0.4080)	5.7213 (0.6011)
	$\alpha = 0, \beta = 1$	2.1677 (0.3047)	3.0974 (0.4432)	2.7247 (0.3535)	2.2970 (0.4642)
	$\alpha = 1, \beta = 0$	5.0248 (0.4583)	6.7614 (0.6470)	5.0542 (0.4011)	6.7858 (0.6075)
	$\alpha = 1, \beta = 1$	3.1567 (0.2814)	4.1654 (0.4399)	3.6296 (0.3351)	4.8813 (0.4526)
$\rho_{T0S0} = \rho_{T1S1} = 0.7$	$\alpha = 0, \beta = 0$	2.6532 (0.4064)	3.7511 (0.6266)	2.6419 (0.4080)	3.7349 (0.6011)
	$\alpha = 0, \beta = 1$	1.2159 (0.3047)	1.6860 (0.4432)	1.6232 (0.3535)	2.2970 (0.4642)
	$\alpha = 1, \beta = 0$	3.6312 (0.4583)	4.7488 (0.6470)	3.6168 (0.4011)	4.7782 (0.6075)
	$\alpha = 1, \beta = 1$	2.2152 (0.2814)	2.7624 (0.4399)	2.6145 (0.3351)	3.2596 (0.4526)
$\rho_{T0S0} = \rho_{T1S1} = 0.9$	$\alpha = 0, \beta = 0$	1.2714 (0.4064)	1.8041 (0.6266)	1.2857 (0.4080)	1.8054 (0.6011)
	$\alpha = 0, \beta = 1$	0.1449 (0.3037)	0.2045 (0.4432)	0.4749 (0.3535)	0.6597 (0.4642)
	$\alpha = 1, \beta = 0$	2.2808 (0.4583)	2.8128 (0.6470)	2.2756 (0.4011)	2.8026 (0.6075)
	$\alpha = 1, \beta = 1$	1.1386 (0.2814)	1.1983 (0.4399)	1.4686 (0.3351)	1.6680 (0.4526)

Note: Mean (SD) of $ISTE_M$ values in the different scenarios.

8 | DISCUSSION

In the meta-analytic framework, Burzykowski and Buyse[2] defined the STE as the minimum value of the expected causal treatment effect on S (i.e., α) that leads to a significant positive expected causal treatment effect on T (i.e., β). The aim of the current paper was to introduce a similar concept in a causal-inference single-trial setting for normally distributed endpoints. The ISTE corresponds to the minimum value of ΔS_0 for which the lower limit of the PI around ΔT exceeds zero. The ISTE and the STE differ (1) in terms of the level at which they operate (ISTE: level of individual patients, STE: level of the clinical trial), and (2) in terms of the data requirements (ISTE: one clinical trial is sufficient, STE: multiple clinical trials are needed), but they are similar in the sense that they supplement metrics of surrogacy that are based on the strength of the association between (individual or expected) causal treatment effects on S and T with a more clinically interpretable metric of surrogacy.

Further, the connections between ISTE and the average causal necessity and sufficiency concepts were discussed. In fact, the sensitivity-based approach for ISTE that was proposed in section 5 can also be used (1) to informally evaluate the plausibility of the average causal necessity assumption in a clinical trial, and (2) to obtain the range of w values for which it holds that $E(\Delta T | \Delta S_0 > w) = 0$ (as was illustrated in the case study, see Figure 4A,B).

Some critical comments and suggestions for future research can be given. First, a Monte Carlo procedure was proposed to address the identifiability problems that occur in the estimation of the ISTE. It is important to point out that this Monte Carlo procedure is not designed to estimate the unidentifiable ISTE, but rather to offer a sensitivity analysis. Basically, the vector of ISTE values that is obtained from the procedure assesses the ISTE across 'plausible realities,' that is, across realities that are compatible with the data at hand. Similar ideas were proposed by References 26 and 27, who used the 'region of ignorance' and 'uncertainty region' concepts in a missing data context. The region of ignorance corresponds to the range of the point estimates for the metric of interest that is obtained under different plausible missing data mechanisms. The uncertainty region includes the sampling uncertainty in addition to the lack of knowledge that is associated with the missing data. Similarly, ISTE can be estimated across different realities that are compatible with the data at hand, with or without accounting for sampling variability. When sampling variability is not accounted for, the sensitivity analysis is conducted by fixing the identifiable quantities to their estimated values. When sampling variability is accounted for (as was done in the case study and the simulation study), the identifiable quantities are sampled from a distribution rather than a fixed value (e.g., a uniform or bootstrapped distribution). Note that the results that were obtained using uniform and bootstrapped distributions in the case study and simulations were very similar, but more extensive simulations are needed to further evaluate the conditions under which both approaches yield comparable results. For example, it is possible that the differences in the results are more substantial when the sample size

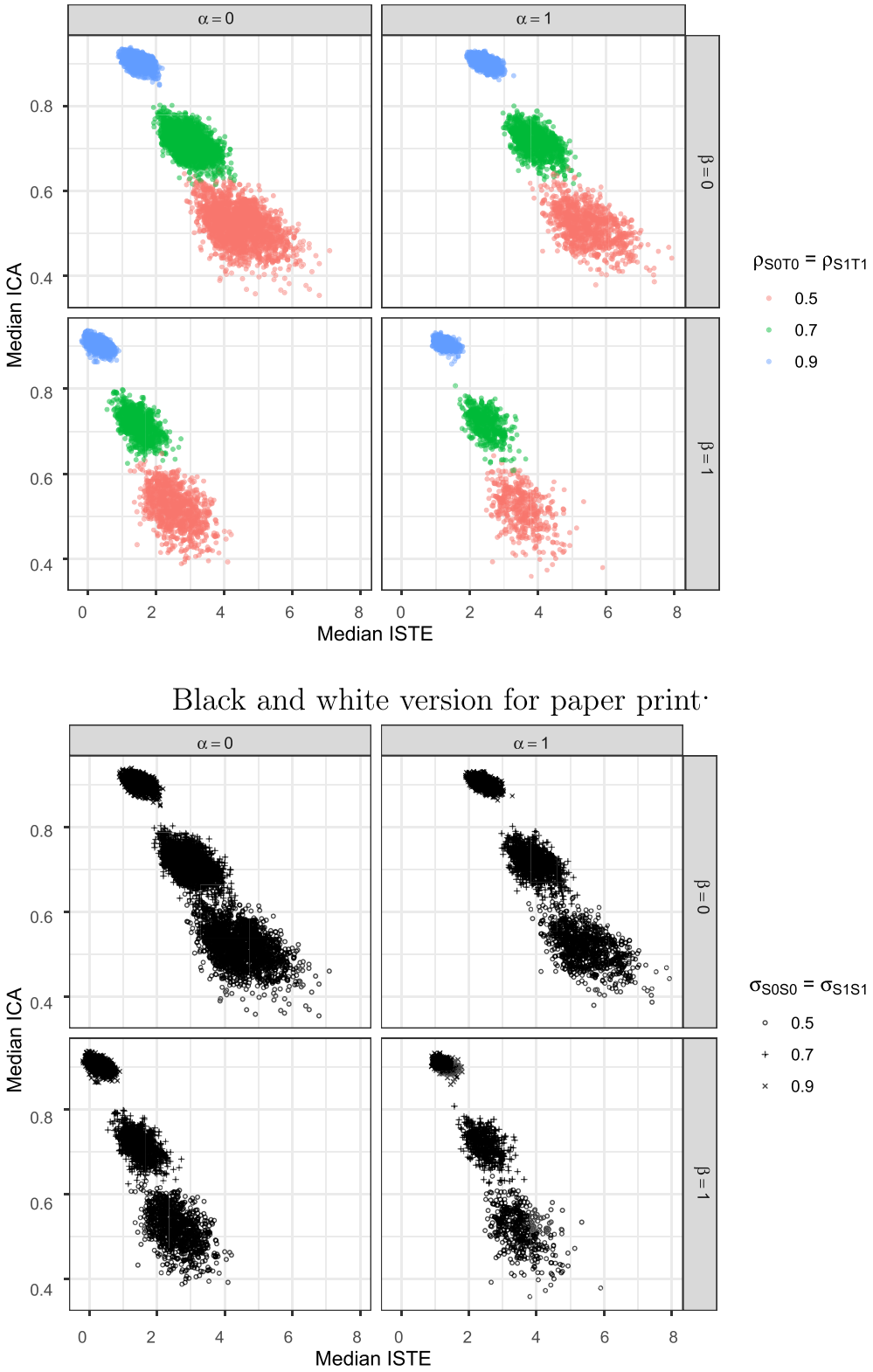

FIGURE 5 Simulation study. Scatter plots of the ICA_M against the $ISTE_M$

is small. The latter issue may, for example, arise when a surrogate endpoint is being evaluated in a rare disease context where the available clinical trial data are sparse.[4]

On a related note, ISTE was estimated across a set of plausible values for the unidentifiable parameters in $\boldsymbol{\Sigma}$ in the sensitivity-based approach. In situations where no biological information is available, the use of $G \sim \text{unif}(-1, 1)$ was

TABLE 2 Coverage in the different simulation scenarios

Identifiable correlations	Expected causal treatment effects	$\sigma_{T0T0} = \sigma_{T1T1} = 1$ $\sigma_{S0S0} = \sigma_{S1S1} = 1$	$\sigma_{T0T0} = \sigma_{T1T1} = 1$ $\sigma_{S0S0} = \sigma_{S1S1} = 2$	$\sigma_{T0T0} = \sigma_{T1T1} = 2$ $\sigma_{S0S0} = \sigma_{S1S1} = 1$	$\sigma_{T0T0} = \sigma_{T1T1} = 2$ $\sigma_{S0S0} = \sigma_{S1S1} = 2$
$\rho_{T0S0} = \rho_{T1S1} = 0.5$	$\alpha = 0, \beta = 0$	98.5%	98.3%	97.5%	98.8%
	$\alpha = 0, \beta = 1$	94.5%	96.5%	95.5%	97.0%
	$\alpha = 1, \beta = 0$	98.0%	98.5%	96.0%	100.0%
	$\alpha = 1, \beta = 1$	94.0%	96.0%	98.0%	98.0%
$\rho_{T0S0} = \rho_{T1S1} = 0.7$	$\alpha = 0, \beta = 0$	96.0%	96.8%	98.3%	98.8%
	$\alpha = 0, \beta = 1$	98.0%	96.5%	99.0%	97.5%
	$\alpha = 1, \beta = 0$	97.0%	99.0%	98.5%	100.0%
	$\alpha = 1, \beta = 1$	97.0%	98.0%	98.0%	98.0%
$\rho_{T0S0} = \rho_{T1S1} = 0.9$	$\alpha = 0, \beta = 0$	99.0%	98.8%	98.5%	98.0%
	$\alpha = 0, \beta = 1$	99.0%	98.0%	99.5%	98.5%
	$\alpha = 1, \beta = 0$	99.5%	98.5%	97.5%	98.5%
	$\alpha = 1, \beta = 1$	99.0%	97.0%	99.0%	100.0%

proposed based on Laplace's principle of indifference and information-theoretic ideas. The use of Laplace's principle of indifference (or similarly, the use of a uniform prior in a Bayesian context) has been the topic of heated debate in statistics for at least a century.[28] For example, Fisher rejected the use of a uniform prior because if it were used for a different parametrisation, the inference would change as well.[29] A detailed discussion of this topic is beyond the scope of the present paper, but from a more practical perspective the use of $G \sim \text{unif}(-1, 1)$ seems to be a sensible choice because the simulation study (see section 7) showed good coverage of ISTE for $G \sim \text{unif}(-1, 1)$. In situations where there is biological or substantive information available for the unidentified correlations, Laplace's principle of indifference no longer applies and the use of a more narrow grid G can be justifiable. In line with this claim, the simulation study showed that the use of a more narrow G leads to better coverage of ISTE—but this is only the case when the assumptions that are made are in fact correct. When the wrong assumptions are made (e.g., $G \sim \text{unif}(0, 1)$ is assumed but the true unidentified correlation is negative), coverage of ISTE is adversely affected. As the biological assumptions that are made in the sensitivity analysis are not verifiable based on the data, caution is thus needed when restricting G.

Second, depending on the disease area and the true endpoint at hand, it is possible that the individual causal treatment effect on T should significantly exceed a threshold value that is higher than 0 to be considered clinically meaningful. In such a scenario, one would be interested in identifying the value ΔS_0 for which it holds that $l(\Delta S_0) = \tau$ (with $\tau =$ some threshold value higher than 0). This value can be obtained in the same way as ISTE (i.e., by setting $l(\Delta S_0)$ equal to τ in Equation (3) and solving the equation for ΔS_0). An example is provided in section 3 of Part I of the web appendix S1, and this procedure is also implemented in the R package surrogate.

Third, surrogate endpoint evaluation methods are typically applied to clinical trial data. In such studies, the identifiability conditions that were described in section 2 are expected to hold 'by design.' For example, the (unconditional) exchangability assumption is plausible by virtue of the randomisation procedure that is used to allocate treatments to patients. As a result, a simple linear regression model like (2) is appropriate to study the relation between ΔS and ΔT, and identifiable parameters like $\alpha = E(\Delta(S))$ that are needed to compute ISTE can be estimated consistently. This is however not the case in observational studies, where baseline confounders might be present (e.g., the distribution of prognostic factors for S and/or T may be different in the treated and untreated groups). In such settings, expert knowledge is needed such that exchangeability can be achieved *conditional* on these confounders. In the latter scenario, the simple linear regression model (2) has to be extended to a mediation-type of model that accounts for the relevant confounders. Such an approach is beyond the scope of the present paper (as it is assumed here that data were collected in a parallel group randomised clinical trial), but extensions of the causal-inference framework to settings where one or more confounders are present have been developed by other authors.[30]

Fourth, missing data issues frequently arise in a surrogate evaluation setting (i.e., the measurement of T is often 'difficult' in some way, otherwise there would be no need for a surrogate), and the question rises how this can be dealt with in practice. A possible strategy was outlined by References 19 and 31, who used multiple imputation (MI) to impute the missing S and T in the dataset at hand. Next, Rubin's rules can be applied to the multiply imputed datasets

to obtain the identifiable quantities that are needed to estimate ISTE (e.g., the variances and covariances between S and T in the two treatment conditions). The latter estimates are then passed to the algorithm that carries out the sensitivity analysis. Notice that MI requires the assumption of a missing at random (MAR) mechanism, that is, the probability of an observation being missing should be conditionally independent of the unobserved outcomes given the observed outcomes.[26] Thus any known and measured covariates or outcomes that are related to the missingness should be included in the imputation model (together with S, T and Z). Observe that the MAR and conditional exchangability assumptions are related in the sense (1) that both assume that there are no unmeasured confounders, and (2) that both are unverifiable assumptions that cannot be demonstrated empirically but rather have to be justified based on substantive knowledge.[26,32]

It should be emphasised that the MI procedure is not intended to deal with the 'missingness' of the potential outcomes themselves.[19] Indeed, potential outcomes like T_0 and T_1 are never simultaneously observed and consequently the data at hand do not contain any information with respect to the unidentifiable correlation ρ_{T0T1}—and the same holds for all other unidentifiable quantities in the causal-inference model. Therefore, any information about the unidentified parameters in the imputed data sets comes from the imputation model itself. So basically one would need to impute the data using several imputation models that assume different values for the unidentified quantities to account for this.[19] Such an approach would be similar to the sensitivity analysis that was used in the current paper.

Finally, the methodology that is described in the current paper is based on the multivariate normality assumption for the potential outcomes, which can only partially be evaluated based on the observable data (i.e., normality can only be evaluated for the distributions of S and T in both treatment groups and for the joint distributions of S and T in both treatment groups). The question may rise to what extent the results are impacted by violations of the unverifiable multivariate normality assumption. To this end, an additional simulation study was conducted in which data were generated (1) using a multivariate uniform distribution (based on the method proposed by Reference 33) and (2) using a multivariate normal distribution. It was of main interest to compare the coverage rates of the [minISTE, maxISTE] intervals in the simulation settings where the multivariate normality assumption is valid or not. Only a summary of the results is given here, more details on the simulation study are available in the web appendix S1 (see section 3 of Part II).

Overall, the results showed that coverage was only marginally lower when the multivariate normality assumption was invalid, with coverage rates equal to 95.75% and 97.23% in the settings where the multivariate normality assumption was invalid and valid, respectively. However, in situations where the identifiable correlations were close to zero, the impact of erroneously assuming multivariate normality was larger. For example, when $\rho_{T0S0} = \rho_{T1S1} = 0$, the coverage rates were 91.3% and 96.7% in the scenarios where the multivariate normality assumption was invalid and valid, respectively. The results thus indicate that the methodology is quite robust against violations of the multivariate normality assumption (unless $\rho_{T0S0} = \rho_{T1S1}$ is low), but more suitable methods to deal with non-normally distributed data have been developed (based on e.g., using Gaussian copula models, see References 2 and 34). The use of such approaches is recommended when the multivariate normality assumption is implausible.

SUPPLEMENTARY MATERIALS

The methodology that is proposed in this manuscript is implemented in the R package surrogate. A web appendix S1 that details the analysis of the case study using this package is available. The web appendix S1 also contains some additional simulation results.

DATA AVAILABILITY STATEMENT

Data are available in R package Surrogate which is available for download at CRAN (see https://CRAN.R-project.org/package=Surrogate).

ORCID

Wim Van der Elst https://orcid.org/0000-0003-4315-7406
Paul Meyvisch https://orcid.org/0000-0002-7248-4273

REFERENCES

1. Alonso AA, 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.

2. Burzykowski T, Molenberghs G, Buyse M. *The Evaluation of Surrogate Endpoints*. New York, NY: Springer-Verlag; 2005.
3. Burzykowski T, Buyse M. Surrogate threshold effect: an alternative measure for meta-analytic surrogate endpoint validation. *Pharm Stat*. 2006;5:173-186.
4. Alonso AA, Bigirumurame T, Burzykowski T, et al. *Applied Surrogate Endpoint Evaluation Methods with SAS and R*. New York, NY: CRC Press; 2016.
5. Buyse M, Molenberghs G. The validation of surrogate endpoints in randomized experiments. *Biometrics*. 1998;54:1014-1029.
6. 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.
7. Freedman LS, Graubard BI, Schatzkin A. Statistical validation of intermediate endpoints for chronic diseases. *Stat Med*. 1992;11:167-178.
8. Prentice RL. Surrogate endpoints in clinical trials: definitions and operational criteria. *Stat Med*. 1989;8:431-440.
9. Elliott MR, Li Y, Taylor JMG. Accommodating missingness when assessing surrogacy via principal stratification. *Clin Trials*. 2013;10:363-377.
10. Van der Elst W, Alonso AA, Geys H, et al. Univariate versus multivariate surrogate endpoints in the single-trial setting. *Stat Biopharm Res*. 2019;11:301-310.
11. Rubin DB. Statistics and causal inference: comment: which ifs have causal answers. *J Am Stat Assoc*. 1986;81:961-962.
12. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70:41-55.
13. Fitzmaurice G, Davidian M, Verbeke G, Molenberghs G. *Longitudinal Data Analysis*. New York, NY: CRC Press; 2009.
14. Conlon ASC, Taylor JMG, Elliott MR. Surrogacy assessment using principal stratification when surrogate and outcome measures are multivariate normal. *Biostatistics*. 2013;14:1-18.
15. Alonso AA, Van der Elst W, Molenberghs G. A maximum entropy approach for the evaluation of surrogate endpoints based on causal inference. *Stat Med*. 2018;37:4525-4538.
16. Wolfinger RD. Tolerance intervals for variance components models using Bayesian simulation. *J Qual Technol*. 1998;30:18-32.
17. Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics*. 2002;58:21-29.
18. Gilbert PB, Hudgens MG. Evaluating candidate principal surrogate endpoints. *Biometrics*. 2008;64:1146-1154.
19. Alonso AA, Van der Elst W, Meyvisch P. Assessing a surrogate predictive function: a causal inference approach. *Stat Med*. 2017;36:1083-1098.
20. Van der Elst W, Molenberghs G, Alonso AA. Exploring the relationship between the causal-inference and meta-analytic paradigms for the evaluation of surrogate endpoints. *Stat Med*. 2016;35:1281-1298.
21. Meyvisch P, Alonso AA, Van der Elst W, Molenberghs G. Assessing the predictive value of a binary surrogate for a binary true endpoint, based on the minimum probability of a prediction error. *Pharm Stat*. 2019;18:304-315.
22. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Diseases*. Washington, DC: American Psychiatric Association; 2000.
23. Overall J, Gorham D. The brief psychiatric rating scale. *Psychol Rep*. 1962;10:799-812.
24. Singh M, Kay S. A comparative study of haloperidol and chlorpromazine in terms of clinical effects and therapeutic reversal with benztropine in schizophrenia. Theoretical implications for potency differences among neuroleptics. *Psychopharmacologia*. 1975;43:103-113.
25. Mortimer AM. Symptom rating scales and outcome in schizophrenia. *Br J Psychiatry*. 2007;191:s7-s14.
26. Molenberghs G, Fitzmaurice G, Kenward MG, Tsiatis A, Verbeke G. *Handbook of Missing Data Methodology*. New York, NY: CRC Press; 2015.
27. Vansteelandt S, Goetghebeur E, Kenward MG, Molenberghs G. Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Stat Sin*. 2006;16:953-979.
28. Gillies DA. Was bayes a Bayesian? *Hist Math*. 1987;14:325-346.
29. Fisher RA. On the mathematical foundations of theoretical statistics. *Philos Trans Royal Soc A*. 1922;222:309-368.
30. VanderWeele TJ, Vansteelandt S. Mediation analysis with multiple mediators. *Epidemiol Methods*. 2014;2:95-115.
31. Van der Elst W. *Statistical Evaluation Methodology for Surrogate Endpoints in Clinical Studies*. [Unpublished PhD dissertation]. 2016.
32. Hernón MA. Beyond exchangeability: the other conditions for causal inference in medical research. *Stat Methods Med Res*. 2011;21:3-5.
33. Falk M. A simple approach to the generation of uniformly distributed random variables with prescribed correlations. *Commun Stat Simul Comput*. 1999;28:785-791.
34. Taylor JMG, Conlon ASC, Elliott MR. Surrogacy assessment using principal stratification with multivariate normal and Gaussian copula models. *Clin Trials*. 2015;12:317-322.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Van der Elst W, Abad AA, Coppenolle H, Meyvisch P, Molenberghs G. The individual-level surrogate threshold effect in a causal-inference setting with normally distributed endpoints. *Pharmaceutical Statistics*. 2021;20(6):1216–1231. https://doi.org/10.1002/pst.2141