Multiple Surrogates in the Meta-Analytic Setting for Normally Distributed Endpoints

Peer-reviewed author version

VAN DER ELST, Wim; ONG, Fenny; Stijven, Florian; Abad, Ariel Alonso; VAN KEILEGOM, Ingrid; GEYS, Helena; Eisele, Lewin & MOLENBERGHS, Geert (2024)
Multiple Surrogates in the Meta-Analytic Setting for Normally Distributed Endpoints.
In: Statistics in biopharmaceutical research,.

DOI: 10.1080/19466315.2024.2429387
Handle: http://hdl.handle.net/1942/44999

Multiple surrogates in the meta-analytic setting for normally distributed endpoints

Wim Van der Elst
Johnson & Johnson, Innovative Medicine

Fenny Ong
I-BioStat, UHasselt

Florian Stijven
I-BioStat, KU Leuven

Ariel Alonso Abad
I-BioStat, KU Leuven

Ingrid Van Keilegom
ORSTAT, KU Leuven

Helena Geys
Johnson & Johnson, Innovative Medicine

Lewin Eisele
Johnson & Johnson, Innovative Medicine

and

Geert Molenberghs
I-BioStat, KU Leuven & UHasselt

Abstract

The identification of good surrogate endpoints is a challenging endeavour. This may, at least partially, be attributable to the fact that most researchers have focused on the identification of a single surrogate endpoint. It is thus implicitly assumed that the treatment effect on the true endpoint (T) can be accurately predicted based on the treatment effect on *one* surrogate endpoint (S) only. Given the complex nature of many diseases and the different therapeutic pathways in which a treatment can impact T, this assumption may be too optimistic. For example, in oncology, the effect of a treatment often depends on *both* the treatment's efficacy and its toxicity.

In the present paper, the meta-analytic framework of Buyse *et al.* (2000) is extended to the setting where multiple S are considered. To cope with potential model convergence issues that often arise in a meta-analytic framework, several simplified model fitting strategies are proposed. Further, simulation studies are conducted to evaluate the properties of the estimated surrogacy metrics, and the new methodology is applied on a case study in schizophrenia. An online Appendix that details how the analyses can be conducted in practice (using the R package *Surrogate*) is also provided.

Word count: 197.

Keywords: Meta-analytic framework; Multiple surrogate endpoints; two-stage approach; individual-level surrogacy; trial-level surrogacy

1 Introduction

The duration, complexity, and cost of a clinical trial are substantially affected by the endpoints that are used to assess treatment efficacy (Burzykowski *et al.*, 2005). 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 with 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 termed a surrogate endpoint (Buyse and Molenberghs, 1998; Buyse *et al.*, 2000; Freedman, Graubard and Schatzkin, 1992; Prentice, 1989).

In spite of important methodological advances in recent years, the identification of good surrogate endpoints remains challenging (Alonso *et al.*, 2016; Buyse *et al.*, 2000, 2010). This may be attributable in part to the fact that most researchers have focused on the identification of a *single* surrogate endpoint. Indeed, it has often implicitly been assumed that the treatment effect on the true endpoint can be accurately predicted based on the treatment effect on *one* surrogate endpoint only. Given the complex nature of many diseases and the various therapeutic pathways in which a treatment can impact the true endpoint, this assumption may be overly optimistic. For example, in oncology, the effect of a treatment often depends on its efficacy and its toxicity (Xu and Zeger, 2001). Having multiple surrogate endpoints that capture the impact of the treatment on both processes can be expected to result in a better prediction of the treatment effect on the true endpoint. Similarly, in neurodegenerative conditions such as Alzheimer's disease, it may be unrealistic to assume that the treatment effect on the true endpoint (which is typically a complex and multifaceted rating scale or cognitive test) can be accurately predicted by a single surrogate endpoint. For example, tau pathology biomarkers in different areas of the brain (e.g., the hippocampus, amygdala, and frontal gyri) have been shown to be only moderately associated with cognition when evaluated individually (with correlations typically below 0.50; Bejanin *et al.*, 2017) – but *taken together*, the different tau pathology measurements

may provide a good multiple surrogate.

In view of these considerations, several authors have argued in favour of using *multiple* surrogate endpoints. For example, Xu and Zeger (2001) proposed a joint model for a time-to-event true endpoint and multiple surrogate endpoints in the single-trial setting (i.e., assuming that the data from only one clinical trial are available). Parast *et al.* (2021) proposed a procedure to quantify surrogacy in the setting where multiple surrogates are available for a time-to-event true endpoint in the single-trial setting based on a dimension-reduction approach. Van der Elst *et al.* (2019) proposed an evaluation strategy for multiple continuous normally distributed surrogates in the single-trial setting based on causal-inference and information-theoretic ideas. Alonso *et al.* (2004) developed a procedure to assess surrogacy for a continuous normally distributed endpoint that is repeatedly measured over time in the meta-analytic setting (i.e., assuming that the data from multiple clinical trials are available) based on canonical correlations.

In the current paper, the gold-standard meta-analytic surrogate evaluation approach of Buyse *et al.* (2000) – that allows for the consideration of one surrogate endpoint only – is extended to the multiple surrogate setting. The remainder of this paper is organised as follows. In Sections 2 and 3, the meta-analytic approach for multiple surrogate endpoints is introduced together with simplified model fitting strategies that can cope with potential model convergence issues. In Section 4, the newly developed methodology is exemplified in a case study. In Section 5, a simulation study is conducted to evaluate model convergence and examine the bias, efficiency and coverage of confidence intervals for the surrogacy metrics. Finally, a critical appraisal regarding the newly proposed methodology is given in Section 6.

Applied statisticians and researchers often encounter challenges with the evaluation of surrogate endpoints due to the lack of user-friendly software. To address this issue, the multiple surrogate evaluation methodology that is proposed in the current paper has been incorporated into an R package *Surrogate* (available for download at CRAN; `https://CRAN.R-project.org/package=Surrogate`). In an online Appendix that accompanies this paper, it is shown how the package can be used to conduct the surrogacy analyses in practice.

2 The meta-analytic approach

Buyse *et al.* (2000) proposed a meta-analytic surrogate evaluation approach in the single surrogate setting. In the current section, the methodology of these authors will be extended to the setting where multiple surrogate endpoints are considered simultaneously.

Let us assume that the data of $i = 1, 2, \ldots, N$ clinical trials are available, in the ith of which $j = 1, 2, \ldots, n_i$ patients are enrolled. Denote by T_{ij} a normally distributed true endpoint T for patient j in trial i, by $\boldsymbol{S} = (S_{1ij},\ S_{2ij}, ...,\ S_{Kij})$ the vector of K normally distributed surrogate endpoints, and by Z_{ij} the (binary) indicator variable for the treatment. It is further assumed that only two treatments are under evaluation ($Z = 0, 1$) in a parallel study design.

In this setting, surrogacy can be evaluated based on the following linear mixed-effects model:

$$\begin{cases} S_{1ij} &= \mu_{S_1} + m_{S_{1i}} + (\alpha_{S_1} + a_{S_{1i}})\, Z_{ij} + \varepsilon_{S_{1ij}}, \\ S_{2ij} &= \mu_{S_2} + m_{S_{2i}} + (\alpha_{S_2} + a_{S_{2i}})\, Z_{ij} + \varepsilon_{S_{2ij}}, \\ &\vdots \\ S_{Kij} &= \mu_{S_K} + m_{S_{Ki}} + (\alpha_{S_K} + a_{S_{Ki}})\, Z_{ij} + \varepsilon_{S_{Kij}}, \\ T_{ij} &= \mu_T + m_{T_i} + (\beta_T + b_{T_i})\, Z_{ij} + \varepsilon_{T_{ij}}, \end{cases} \tag{1}$$

where $\mu_{S_1},\ \mu_{S_2},\ \ldots,\ \mu_{S_K}$ and μ_T are the fixed intercepts for $S_1,\ S_2,\ \ldots,\ S_K$ and T, $m_{S_{1i}},\ m_{S_{2i}},\ \ldots,\ m_{S_{Ki}}$, and m_{T_i} are the corresponding random intercepts, $\alpha_{S_1},\ \alpha_{S_2},\ \ldots,$ α_{S_K} and β_T are the fixed treatment effects for $S_1,\ S_2,\ \ldots,\ S_K$ and T, and $a_{S_{1i}},\ a_{S_{2i}},$ $\ldots,\ a_{S_{Ki}}$ and b_{T_i} are the corresponding random treatment effects. The vectors of the random effects $(m_{S_{1i}},\ m_{S_{2i}}, \ldots,\ m_{S_{Ki}},\ m_{T_i},\ a_{S_{1i}},\ a_{S_{2i}}, \ldots,\ a_{S_{Ki}},\ b_{T_i})$ and the error terms $\left(\varepsilon_{S_{1ij}},\ \varepsilon_{S_{2ij}},\ ...,\ \varepsilon_{S_{Kij}},\ \varepsilon_{T_{ij}}\right)$ are assumed to be mean-zero normally distributed with un-

structured variance-covariance matrices $\mathbf{D}$ and $\mathbf{\Sigma}$, respectively:

$$\mathbf{D} = \begin{pmatrix} d_{m_{S_1},m_{S_1}} & d_{m_{S_1},m_{S_2}} & \dots & d_{m_{S_1},m_{S_K}} & d_{m_{S_1},m_T} & d_{m_{S_1},a_{S_1}} & d_{m_{S_1},a_{S_2}} & \dots & d_{m_{S_1},a_{S_K}} & d_{m_{S_1},b_T} \\ & d_{m_{S_2},m_{S_2}} & \dots & d_{m_{S_2},m_{S_K}} & d_{m_{S_2},m_T} & d_{m_{S_2},a_{S_1}} & d_{m_{S_2},a_{S_2}} & \dots & d_{m_{S_2},a_{S_K}} & d_{m_{S_2},b_T} \\ & & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ & & & d_{m_{S_K},m_{S_K}} & d_{m_{S_K},m_T} & d_{m_{S_K},a_{S_1}} & d_{m_{S_K},a_{S_2}} & \dots & d_{m_{S_K},a_{S_K}} & d_{m_{S_K},b_T} \\ & & & & d_{m_T,m_T} & d_{m_T,a_{S_1}} & d_{m_T,a_{S_2}} & \dots & d_{m_T,a_{S_K}} & d_{m_T,b_T} \\ & & & & & d_{a_{S_1},a_{S_1}} & d_{a_{S_1},a_{S_2}} & \dots & d_{a_{S_1},a_{S_K}} & d_{a_{S_1},b_T} \\ & & & & & & d_{a_{S_2},a_{S_2}} & \dots & d_{a_{S_2},a_{S_K}} & d_{a_{S_2},b_T} \\ & & & & & & & \ddots & \vdots & \vdots \\ & & & & & & & & d_{a_{S_K},a_{S_K}} & d_{a_{S_K},b_T} \\ & & & & & & & & & d_{b_T,b_T} \end{pmatrix}, \quad (2)$$

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{S_1,S_1} & \sigma_{S_1,S_2} & \dots & \sigma_{S_1,S_K} & \sigma_{S_1,T} \\ & \sigma_{S_2,S_2} & \dots & \sigma_{S_2,S_K} & \sigma_{S_2,T} \\ & & \ddots & \vdots & \vdots \\ & & & \sigma_{S_K,S_K} & \sigma_{S_K,T} \\ & & & & \sigma_{T,T} \end{pmatrix}. \quad (3)$$

In the meta-analytic framework, surrogacy is quantified by two metrics, i.e., the trial- and individual-level coefficients of determination. The trial-level coefficient of determination quantifies the strength of the association between the trial-specific treatment effects on T and the trial-specific intercepts and treatment effects on $\boldsymbol{S}$ in the N different trials:

$$R^2_{trial(f)} = R^2_{b_{T_i}|m_{S_{1i}}, m_{S_{2i}}, ..., m_{S_{Ki}}, a_{S_{1i}}, a_{S_{2i}}, \dots a_{S_{Ki}}} = \frac{\boldsymbol{D}^T_{ST}\, \boldsymbol{D}^{-1}_{SS}\, \boldsymbol{D}_{ST}}{d_{b_T,b_T}}. \quad (4)$$

All quantities in (4) are based on the $\mathbf{D}$ matrix, with $\mathbf{D}_{SS}$ corresponding to the variance-covariance matrix of the random intercepts and treatment effects for $\boldsymbol{S}$, and $\mathbf{D}_{ST}$ corresponding to the column vector of the covariances between the random intercepts and treatment effects for $\boldsymbol{S}$ and the random treatment effect for T.

The R^2_{trial} value is unitless and lies within the unit interval when the $\mathbf{D}$ matrix is positive-definite. An exact 95% confidence interval (CI) around R^2_{trial} can be obtained using the procedure that was proposed by Lee (1971). Alternatively, a non-parametric bootstrap or the Delta method can be used to get approximate CIs (Cortiñas *et al.*, 2008). An R^2_{trial}

that is close to 1 (taking its CI into account) indicates that there is a strong association between the treatment effects on $\boldsymbol{S}$ and T across the N different trials. A surrogate is called trial-level valid when this is the case. The term "trial-level" surrogacy refers to the fact that the treatment effects on $\boldsymbol{S}$ and T are estimated at the level of the clinical trials. Note that the (f) indicator in the $R^2_{trial(f)}$ subscript is used to indicate that a so-called full model is used to evaluate surrogacy – as opposed to the situation where a reduced model is used (see Section 3.2 below).

The individual-level coefficient of determination quantifies the strength of the association between $\boldsymbol{S}$ and T at the level of the individual patients (after adjustment for both the trial- and treatment-effects):

$$R^2_{indiv} = R^2_{\varepsilon_{T_{ij}}|\varepsilon_{S_{1ij}},\varepsilon_{S_{2ij}},...,\varepsilon_{S_{Kij}}} = \frac{\boldsymbol{\Sigma}_{ST}^T \, \boldsymbol{\Sigma}_{SS}^{-1} \, \boldsymbol{\Sigma}_{ST}}{\sigma_{T,T}}. \tag{5}$$

All quantities in (5) are based on the $\boldsymbol{\Sigma}$ matrix, with $\boldsymbol{\Sigma}_{SS}$ corresponding to the variance-covariance matrix of the errors for $\boldsymbol{S}$, and $\boldsymbol{\Sigma}_{ST}$ corresponding to the column vector of the covariances between the errors for $\boldsymbol{S}$ and T.

An exact 95% CI around R^2_{indiv} can again be obtained using the approach of Lee (1971), or alternatively a non-parametric bootstrap or the Delta method can be used to get approximate CIs. An R^2_{indiv} that is close to 1 (taking its CI into account) indicates that there is a strong association between $\boldsymbol{S}$ and T at the level of the individual patients (after accounting for treatment- and trial-effects). A surrogate is called individual-level valid when this is the case. The term "individual-level" surrogacy refers to the fact that the $\boldsymbol{S}$ and T are measured at the level of the individual patients. In essence, $\boldsymbol{S}$ is individual-level valid when it has a good prognostic value for T (Buyse *et al.*, 2022).

3 Simplified modelling strategies

The mixed-effects modelling approach that was detailed in Section 2 poses considerable computational challenges. Indeed, fitting a linear mixed-effects model is typically done using Newton-Raphson or quasi-Newton optimisation methods (for details, see Lindstrom and Bates, 1988; Verbeke and Molenberghs, 2000). Based on some starting values for the parameters at hand, these procedures iteratively update the parameter estimates until the

convergence criteria are met. Unfortunately, the optimization methods may not converge when complex linear mixed-effects models are considered. This means that the iterative process does not converge at all, or that it converges to values that are close to or outside the boundary of the parameter space (i.e., variances that are close to zero or even negative). Simulation studies in the single surrogate setting have shown that such problems mainly occur (i) when the number of trials is small, (ii) when the size of the between-trial variability (i.e., the components in the $\mathbf{D}$ matrix is small relative to the size of the residual variability (i.e., the components in the $\mathbf{\Sigma}$ matrix, and (iii) when the number of patients in the different trials is unbalanced (Buyse *et al.*, 2000; Burzykowski *et al.*, 2005; Renard *et al.*, 2002; Ong *et al.*, 2022; Van der Elst *et al.*, 2015).

Unfortunately, the conditions that are described in (i)–(iii) are often encountered in real-life surrogate evaluation settings, and thus convergence problems are prevalent. Such issues can be expected to be even exacerbated in the current setting where multiple surrogates are of interest, because the consideration of each additional surrogate increases the complexity of the model. Buyse *et al.* (2000) and Tibaldi *et al.* (2003) have proposed a number of simplified model fitting strategies that can be used when model convergence problems occur in the single surrogate setting. In particular, these authors have proposed to simplify model (1) along four dimensions. Here, these simplified model fitting strategies are generalized to the multiple surrogate endpoint evaluation setting.

3.1 The trial dimension: fixed- versus random-effects models

To avoid the computational problems that arise in the estimation of the variance components of model (1), the mixed-effects model can be replaced by its fixed-effects (i.e., two-stage) counterpart (Buyse *et al.*, 2000; Tibaldi *et al.*, 2003). When the fixed-effects approach is used, either a single multivariate or multiple univariate linear regression models are fitted to the data of each of the N trials separately. The choice for a multivariate or a univariate modelling approach is determined by the assumptions that are made regarding the association structure of the errors (see Section 3.3). Thus, in Stage 1, the following

(multi- or univariate) fixed-effects models are fitted:

$$\begin{cases} S_{1ij} &= \mu_{S_{1i}} + \alpha_{S_{1i}} Z_{ij} + \varepsilon_{S_{1ij}}, \\ S_{2ij} &= \mu_{S_{2i}} + \alpha_{S_{2i}} Z_{ij} + \varepsilon_{S_{2ij}}, \\ &\vdots \\ S_{Kij} &= \mu_{S_{Ki}} + \alpha_{S_{Ki}} Z_{ij} + \varepsilon_{S_{Kij}}, \\ T_{ij} &= \mu_{Ti} + \beta_{T_i} Z_{ij} + \varepsilon_{T_{ij}}, \end{cases} \tag{6}$$

where $\mu_{S_{1i}}$, $\mu_{S_{2i}}, \ldots, \mu_{S_{Ki}}$, and μ_{Ti} are the trial-specific intercepts on $\boldsymbol{S}$ and T, and $\alpha_{S_{1i}}, \alpha_{S_{2i}}, \ldots, \alpha_{S_{Ki}}$ and β_{T_i} are the corresponding trial-specific treatment effects. The error terms $\varepsilon_{S_{1ij}}$, $\varepsilon_{S_{2ij}}, \ldots, \varepsilon_{S_{Kij}}$, and $\varepsilon_{T_{ij}}$ are assumed to be mean-zero normally distributed with variance-covariance matrix $\boldsymbol{\Sigma}$ when a multivariate regression model is fitted, or are assumed to be independent when multiple univariate models are fitted (see Section 3.3). The fixed-effects parameter estimates for $\mu_{S_{1i}}$, $\alpha_{S_{1i}}$, $\mu_{S_{2i}}$, $\alpha_{S_{2i}}$, $\ldots$, $\mu_{S_{Ki}}$, $\alpha_{S_{Ki}}$, and β_{T_i} that are obtained by fitting model (6) are subsequently used in Stage 2 of the analysis, where the following multiple linear regression model is fitted:

$$\widehat{\beta}_{T_i} = \lambda_0 + \lambda_1 \widehat{\mu}_{S_{1i}} + \lambda_2 \widehat{\alpha}_{S_{1i}} + \lambda_3 \widehat{\mu}_{S_{2i}} + \lambda_4 \widehat{\alpha}_{S_{2i}} + \cdots + \lambda_{2K-1} \widehat{\mu}_{S_{Ki}} + \lambda_{2K} \widehat{\alpha}_{S_{Ki}} + \varepsilon_i. \tag{7}$$

The classical coefficient of determination that is obtained by fitting model (7) provides an estimate for $R^2_{trial(f)}$. Similarly to what was the case when the mixed-effects model is used to estimate trial-level surrogacy (see expression (4)), the (f) indicator in the $R^2_{trial(f)}$ subscript indicates that a full model is used. Alternatively, trial-level surrogacy can be estimated based on a reduced mixed- or fixed-effects modelling approach (see the next section).

3.2 The model dimension: full versus reduced models

Model (1) is referred to as the *full* mixed-effects model, i.e., the model that contains random intercepts and random treatment effects for $\boldsymbol{S}$ and T. Similarly, model (6) is referred to as the full (multi- or univariate) fixed-effects model, i.e., the model that contains both trial-specific intercepts and treatment effects for $\boldsymbol{S}$ and T.

The random-effects structure of the full multivariate mixed-effects model (1) can be simplified by assuming that there is no heterogeneity in the random intercepts for $\boldsymbol{S}$ and T:

$$\begin{cases} S_{1ij} &= \mu_{S_1} + (\alpha_{S_1} + a_{S_{1i}})\, Z_{ij} + \varepsilon_{S_{1ij}}, \\ S_{2ij} &= \mu_{S_2} + (\alpha_{S_2} + a_{S_{2i}})\, Z_{ij} + \varepsilon_{S_{2ij}}, \\ \vdots & \\ S_{Kij} &= \mu_{S_K} + (\alpha_{S_K} + a_{S_{Ki}})\, Z_{ij} + \varepsilon_{S_{Kij}}, \\ T_{ij} &= \mu_T + (\beta_T + b_{T_i})\, Z_{ij} + \varepsilon_{T_{ij}}. \end{cases}$$

As can be seen, the trial-specific intercepts in model (1) are now replaced by common intercepts, and the $\mathbf{D}$ matrix simplifies accordingly. The $\mathbf{D}_{SS}$ and $\mathbf{D}_{ST}$ components now correspond to the variance-covariance matrix of the random treatment effects for $\boldsymbol{S}$, and the vector of the covariances between the random treatment effects for $\boldsymbol{S}$ and the random treatment effect for T, respectively. These submatrices are referred to as $\mathbf{D}_{SS(r)}$ and $\mathbf{D}_{ST(r)}$, respectively. The (r) subscript indicates that a reduced mixed-effects model is fitted (as opposed to the situation where a full mixed-effect model is used to estimate trial-level surrogacy, see expression (4)). The computation of the trial-level coefficient of determination now simplifies to:

$$R^2_{trial(r)} = R^2_{b_{T_i}|a_{S_{1i}},\, a_{S_{2i}},\, ...,\, a_{S_{Ki}}} = \frac{\boldsymbol{D}^T_{ST(r)}\, \boldsymbol{D}^{-1}_{SS(r)}\, \boldsymbol{D}_{ST(r)}}{d_{b_T, b_T}}.$$

When the fixed-effects approach is used (see Section 3.1), model (6) can be simplified by assuming common intercepts for $\boldsymbol{S}$ and T in Stage 1 of the analysis. Thus, the trial-specific $\mu_{S_{1i}}, \mu_{S_{2i}}, \ldots, \mu_{S_{Ki}}$ and μ_{Ti} intercepts in model (6) are replaced by common intercepts (i.e., $\mu_{S_1}, \mu_{S_2}, \ldots, \mu_{S_K}$ and μ_T, respectively). Further, the $\lambda_1 \widehat{\mu}_{S_{1i}}$, $\lambda_3 \widehat{\mu}_{S_{2i}}$, $\ldots$, and $\lambda_{2K-1} \widehat{\mu}_{S_{Ki}}$ components are dropped from expression (7) in Stage 2 of the analysis.

3.3 The endpoint dimension: univariate versus multivariate models

The error terms for $\boldsymbol{S}$ and T can be assumed to be independent rather than dependent by fitting $K + 1$ univariate models instead of one multivariate (mixed- or fixed-effects)

model. This proposal seems odd at first sight, because it is natural to assume that $\boldsymbol{S}$ and T are correlated in a surrogate endpoint evaluation setting. Nonetheless, making the simplifying assumption that the error terms are uncorrelated is not necessarily a problem. The reason for this is that the explicit consideration of the correlated nature of $\boldsymbol{S}$ and T is mainly of importance to obtain the $\boldsymbol{\Sigma}$ matrix (which is used to estimate R^2_{indiv}, see expression (5)), and often the focus of the analysis is on trial-level surrogacy rather than on individual-level surrogacy. Indeed, it is typically of main interest to examine the extent to which the *treatment effect* on T can be predicted based on the *treatment effects* on $\boldsymbol{S}$ (this is particularly the case for pharmaceutical companies and regulatory agencies). It has been shown that the R^2_{trial} values that are obtained by using univariate or multivariate mixed-effects models are similar in the single surrogate setting (Tibaldi *et al.*, 2003), and the R^2_{trial} values that are obtained by using univariate or multivariate fixed-effects models are identical (Johnson and Wichern, 2007). Moreover, when univariate models are fitted and interest is also in R^2_{indiv}, one can always approximate this quantity by computing the squared multiple correlation between the residuals of $\boldsymbol{S}$ and T (in the same spirit as the multivariate adjusted association; for details, see Van der Elst *et al.*, 2019).

3.4 The measurement error dimension: weighted versus unweighted models

When the (full or reduced) multivariate mixed-effects modelling approach is not used, one is confronted with measurement error because the trial-specific treatment effects on $\boldsymbol{S}$ and T are estimated with error. The magnitude of this error can be assumed to be inversely related to the number of patients in a particular trial. Therefore, a straightforward approach to address this issue is to use a weighted regression model with the trial sizes as the weights in Stage 2 of the analysis (Burzykowski *et al.*, 2005; Tibaldi *et al.*, 2003). Note that the measurement error dimension is not relevant when a multivariate mixed-effects model is used, because it is automatically accounted for and therefore no explicit corrections are needed.

4 A case study in schizophrenia

The methodology will be illustrated in a case study in schizophrenia. Schizophrenia is a psychiatric condition that is hallmarked by hallucinations and delusions (American Psychiatric Association, 2000). The data were collected in five double-blind clinical trials in which the patients were randomly allocated to two treatment arms (i.e., the experimental treatment risperidone versus an active control). A total of $1,941$ patients participated in the five clinical trials, of whom $1,450$ patients received the experimental treatment and 497 patients were given an active control. Details on the different clinical trials can be found in Blin *et al.* (1996), Chouinard *et al.* (1993), Hoyberg *et al.* (1993), Huttunen *et al.* (1995) and Peuskens *et al.* (1995).

In each of the five trials, the Positive and Negative Syndrome Scale (PANSS; Singh and Kay, 1975) was administered. The PANSS is a standardized instrument that is used to rate the symptom severity of people with schizophrenia. It consists of 30 items that are measured on a 7-point Likert scale. These items can be grouped into five factors, i.e., Negative symptoms, Positive symptoms, Cognitive symptoms, Excitement, and Depression (Lindenmayer *et al.*, 1995). The aim of the analysis is to evaluate whether one or more of the 5 PANSS subscales provide a good (multiple) surrogate for T = PANSS Total score.

The data of only five clinical trials were available, which is insufficient to apply the meta-analytic method using trial as the clustering unit (Burzykowski *et al.*, 2005). In the different trials, information was also available regarding the psychiatrists who treated the patients. Treating physician was therefore used as the clustering unit instead of trial in the analyses below. The patients were treated by a total of $N = 126$ psychiatrists. Each of the psychiatrists treated between $n_i = 5$ and 52 patients. Notice that the use of treating physician as the clustering unit can impact the baseline balance of covariates for T and/or $\boldsymbol{S}$, because the randomisation was conducted at the level of the clinical trials (and not at the level of the treating physicians). The estimated treating physician-specific treatment effects can thus be confounded by such imbalances. Along these lines, simulation studies have shown that shifting between clustering units can bias the estimated R^2_{trial} – in particular when the magnitude of the variability of the treatment effects across the different clustering levels varies substantially (see Cortiñas *et al.*, 2004). In the current case study

analysis, it is assumed that no such imbalances for the baseline covariates occur by using treating physician as the clustering unit. For a further discussion on the choice of clustering units and its impact on the metrics of surrogacy, see Chapter 8 of Burzykowski *et al.* (2005).

Table 1 shows the Pearson correlations between the PANSS subscales and the PANSS Total score in the two treatment arms (across the five different clinical trials). As can be seen, the scores on the different subscales are low to moderately intercorrelated with each other (range of Pearson correlations: $[0.291; 0.678]$ and $[0.336; 0.653]$ in the active control and experimental treatment arms, respectively), and more strongly correlated with the PANSS Total score (range $[0.722; 0.805]$ and $[0.693; 0.812]$ in the active control and experimental treatment arms, respectively).
<Insert Table 1 about here>

Results A hierarchical model-building approach with forward selection was used to identify the best $\boldsymbol{S}$. To this end, single surrogacy analyses were first conducted for each of the 5 candidate surrogates separately (i.e., the 5 PANSS subscales), and it was subsequently evaluated whether the consideration of additional surrogates led to an increase in R^2_{trial} and R^2_{indiv}.

When full multivariate mixed-effects models were fitted to evaluate surrogacy (see Section 2), convergence issues emerged for all the models that were considered. As was mentioned in Section 3, such model fitting issues are expected when the full multivariate mixed-effects modelling approach is used. To avoid these computational issues, the mixed-effects model was replaced by its fixed-effects counterpart (i.e., the full multivariate weighted fixed-effects model, see Section 3).

In the single surrogacy analyses, the R^2_{trial} values (and the 95% CIs that were obtained using the procedure proposed by Lee, 1971) ranged between 0.450 (95% CI $[0.308; 0.567]$) for the Depression subscale and 0.658 (95% CI $[0.544; 0.742]$) for the Cognition subscale. The R^2_{indiv} values were of similar magnitude and ranged between 0.487 (95% CI $[0.454; 0.518]$) and 0.652 (95% CI $[0.626; 0.676]$) for the Depression and Cognition subscales, respectively. The single S that had the highest R^2_{trial} and R^2_{indiv} values is thus the Cognition subscale. Table 2 shows the R^2_{trial} and R^2_{indiv} values for all single S, together with the AIC/BIC values

for the fitted Stage 2 models at hand (that are used to estimate R^2_{trial}). As expected, the AIC/BIC values for the Cognition subscale were also the lowest of the 5 single S under consideration.

In the second step, multiple surrogacy analyses were conducted that included the first identified "best" single S (i.e., the Cognition subscale) together with a second S (i.e., the Negative, Positive, Excitement or Depression subscales). Likelihood-ratio (LR) tests were used to compare the fit of the different (nested) Stage 2 models. For example, the best pair of surrogates $\boldsymbol{S}$ that resulted in the highest R^2_{trial} and R^2_{indiv} values consisted of the combination of the Cognition and the Depression subscales, yielding $R^2_{trial} = 0.866$ (95% CI $[0.808; 0.901]$) and $R^2_{indiv} = 0.831$ (95% CI $[0.816; 0.844]$). The LR-test that compares the fitted Stage 2 model that only includes the treating physician-specific intercepts and treatment effects for the Cognition subscale (i.e., the restricted model, with $R^2_{trial} = 0.658$) versus the model that includes the treating physician-specific intercepts and treatment effects for both the Cognition and Depression subscales (i.e., the unrestricted model, with $R^2_{trial} = 0.866$) yielded $\chi^2 = 117.8$ (DF = 2) with p-value < 0.001. Based on the LR-test, the multiple $\boldsymbol{S}$ is thus preferred over the single S. Figure 1 visually illustrates the R^2_{trial} and R^2_{indiv} results for this $\boldsymbol{S}$ graphically (see panels [a] and [b], respectively).

Interestingly, the best pair of multiple surrogates $\boldsymbol{S}$ combines the best and the worse of the single S (i.e., the Cognition and Depression subscales, respectively). So even though the Depression subscale alone is a poor surrogate for the PANSS Total score (with $R^2_{trial} = 0.450$), the combination of the Depression and the Cognition subscales resulted in the best pair of multiple surrogates $\boldsymbol{S}$ (with R^2_{trial} almost doubling to 0.868). Recall from Table 1 that the correlation between the Cognition and Depression subscales was quite low (i.e., 0.405 and 0.339 in both treatment arms), and thus both scales have a relatively small overlap. This low correlation may account for this finding, i.e., both scales appear to capture different aspects of schizophrenia (as opposed to the other PANSS subscales that are more highly inter-correlated with the Cognition subscale – and thus the addition of such subscales adds less information in the surrogacy analysis).

<Insert Figure 1 about here>

The model-building procedure was repeated for 3, 4, and 5 surrogates. As can be seen in Table 2, the best multiple $\boldsymbol{S}$ (as evaluated based on a series of LR-tests) contains all 5 PANSS subscales, with $R^2_{trial} = 0.989$ (95% CI $[0.983; 0.992]$) and $R^2_{indiv} = 0.990$ (95% CI $[0.989; 0.991]$). Both metrics of surrogacy are thus close to 1, which is not surprising because the PANSS Total score is actually the sum of the items of the 5 subscales plus 5 additional items that are not part of these subscales (see Discussion). The high levels of trial- and individual-level surrogacy indicate that the (treatment effects on) the PANSS Total score can be highly accurately predicted based on the (treatment effects) on the PANSS subscales. Figure 2 graphically illustrates the main results. Note that the 95% CIs for R^2_{indiv} are substantially more narrow than those for R^2_{trial}. This is in line with expectations, because R^2_{indiv} is estimated at the level of the individual patients whereas R^2_{trial} is estimated at the level of the treating physicians. The number of patients is much higher than the number of treating physicians (i.e., $1,941$ versus 126), and thus the CIs are narrower.

Table 2 and Figure 2 only summarize the main results. A comprehensive table that shows the estimated R^2_{trial} and R^2_{indiv} for all possible combinations of 1, 2, ..., and 5 surrogates is provided in the online Appendix.

<Insert Table 2 about here>

<Insert Figure 2 about here>

5 Simulation study

A simulation study was conducted which aimed at (i) estimating the convergence rates of the full multivariate mixed-effects modelling approach, and (ii) evaluating the bias, efficiency and coverage of the 95% CIs of the estimated R^2_{trial} and R^2_{indiv} when using the full multivariate mixed- and fixed-effects modelling approaches. The following true data-

generating mechanism was assumed (based on Ong *et al.*, 2022; Van der Elst *et al.*, 2015):

$$\begin{cases} S_{1ij} &= 450 + m_{S_{1i}} + (300 + a_{S_{1i}})\, Z_{ij} + \varepsilon_{S_{1ij}}, \\ S_{2ij} &= 460 + m_{S_{2i}} + (350 + a_{S_{2i}})\, Z_{ij} + \varepsilon_{S_{2ij}}, \\ S_{3ij} &= 470 + m_{S_{3i}} + (400 + a_{S_{3i}})\, Z_{ij} + \varepsilon_{S_{3ij}}, \\ T_{ij} &= 500 + m_{T_i} + (500 + b_{T_i})\, Z_{ij} + \varepsilon_{T_{ij}}, \end{cases}$$

where $(m_{S_{1i}}, m_{S_{2i}}, m_{S_{3i}}, m_{T_i}, a_{S_{1i}}, a_{S_{2i}}, a_{S_{3i}}, b_T) \sim N(\mathbf{0}, \mathbf{D})$ with:

$$\mathbf{D} = \begin{pmatrix} 1000 & 100 & 100 & 100 & 0 & 0 & 0 & 0 \\ 100 & 1000 & 100 & 100 & 0 & 0 & 0 & 0 \\ 100 & 100 & 1000 & 100 & 0 & 0 & 0 & 0 \\ 100 & 100 & 100 & 1000 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1000 & 250 & 250 & 500 \\ 0 & 0 & 0 & 0 & 250 & 1000 & 250 & 500 \\ 0 & 0 & 0 & 0 & 250 & 250 & 1000 & 500 \\ 0 & 0 & 0 & 0 & 500 & 500 & 500 & 1000 \end{pmatrix},$$

and $(\varepsilon_{S_{1ij}}, \varepsilon_{S_{2ij}}, \varepsilon_{S_{3ij}}, \varepsilon_{T_{ij}}) \sim N(\mathbf{0}, \mathbf{\Sigma})$ with:

$$\mathbf{\Sigma} = \begin{pmatrix} 300 & 75 & 75 & 150 \\ 75 & 300 & 75 & 150 \\ 75 & 75 & 300 & 150 \\ 150 & 150 & 150 & 300 \end{pmatrix}.$$

The true R^2_{trial} and R^2_{indiv} both equal 0.5. Datasets were generated that included $N = 10, 20, 50$ trials with $n_i = n = 20$ patients each. Treatment was balanced within a trial. For each N, $1,000$ datasets were generated (so $3,000$ datasets in total). The full multivariate mixed-effects model (1) and its full multivariate fixed-effects counterpart (see Section 3) were fitted to each of the generated datasets. Observe that the measurement error dimension (see Section 3.4) is not relevant here because the number of patients is balanced across the trials. Thus, the use of a weighted or an unweighted Stage 2 model gives the same results.

The simulation study aims to mimic a scenario where an alternative clustering unit is used to assess trial-level surrogacy (e.g., treating physician instead of clinical trial). Indeed, in practice there are often insufficient clinical trials available and in such a scenario one typically obtains many clusters with a relatively small sample size N (as was also the case in the case study; see Section 4).

Convergence The first aim of the simulation study was to examine the extent to which the full multivariate mixed-effects models properly converged. In line with Ong *et al.* (2022), proper convergence was defined as convergence to a positive-definite $\mathbf{D}$ matrix with a condition number below 100. The proper convergence rates for the analyses with $N = 10, 20, 50$ trials equalled 14.2%, 97.5% and 100.0%, respectively. The convergence rate was thus strongly impacted by the number of available trials, with high convergence rates when the data of at least $N = 20$ trials are available (in line with earlier simulations in the single surrogacy setting, see Buyse *et al.*, 2000; Ong *et al.*, 2022; Van der Elst *et al.*, 2015).

Bias, efficiency and coverage The second aim of the simulation study was to examine the bias (i.e., the mean difference between the estimated R^2_{trial} / R^2_{indiv} and their true values), efficiency (i.e., the SD of the estimated values), and coverage (i.e., the probability that the 95% CIs include the true R^2_{trial} / R^2_{indiv} values; the CIs were estimated using the approach of Lee, 1971).

Table 3 summarizes the results. Overall, the results were similar for the full multivariate mixed- and fixed-effects models – except for the bias in R^2_{trial}, which was substantially smaller in the mixed-effects model when $N = 10$ (i.e., bias = 0.220 versus 0.315 for the mixed- and fixed-effects models, respectively). As expected, the bias decreased and the efficiency improved (i.e., lower SDs of the estimated values) when the number of trials increased (in line with earlier simulation studies in the single surrogate setting, see Ong *et al.*, 2022; Van der Elst *et al.*, 2015). The bias remained however non-negligible when $N = 50$. Fortunately, the coverage was close to 95% for all N that were considered. So even though the point estimate is positively biased, the coverage was good even when the number of trials was as small as 10.

For R^2_{indiv}, the bias was small for all N. As expected, the efficiency improved when N

increased. Further, coverage was again close to 95% for all N.

<Insert Table 3 about here>

Comparison with case study results Observe that the convergence rates of the full multivariate mixed-effects models in the simulation study were substantially higher than those in the case study. For example, in the simulation study there was 97.5% convergence when $N = 20$. In contrast, none of the models that were considered in the case study converged (even though N was as large as 126). This is probably attributable to the fact that the data-generating mechanism and the analysis model are in perfect agreement in the simulation setting (unlike in the case study, where the data-generating mechanism is unknown). Moreover, in the simulation study, the between-cluster variability was large (relative to the within-cluster variability) and the number of patients was perfectly balanced across the trials. Previous simulation studies have shown that these conditions ameliorate model convergence issues (Buyse *et al.*, 2000; Ong *et al.*, 2022; Van der Elst *et al.*, 2015).

6 Discussion

At present, most surrogate endpoint evaluation methods allow for considering only one S. Given the complex nature of many diseases and the various therapeutic pathways in which a treatment can impact the clinical outcome, it seems reasonable to assume that the prediction of the treatment effect on T can be substantially improved when multiple surrogates ($\boldsymbol{S}$) are considered rather than only a single one. An example was provided in the case study, where the best single surrogate was only of moderate magnitude (i.e., $R^2_{trial} = 0.658$ and $R^2_{indiv} = 0.652$ for the Cognition subscale), whereas the multiple $\boldsymbol{S}$ that considered all 5 subscales was an almost perfect surrogate at both the trial and the individual levels (i.e., $R^2_{trial} = 0.989$ and $R^2_{indiv} = 0.990$).

The main aim of this paper was to generalize the gold-standard meta-analytic approach to the setting where multiple surrogate endpoints are considered. As expected, model fitting issues were prevalent when the full multivariate mixed-effects approach was used. To overcome these issues, the simplified model fitting strategies that were proposed by Tibaldi *et al.* (2003) in the single surrogate setting were generalized to the multiple surro-

gate setting. The simulation study that was detailed in Section 5 indicated that the full multivariate fixed- and mixed-effects modelling approaches yielded similar results – except for the bias in R^2_{trial}, which was substantially larger in the fixed-effects modelling approach when N was small.

Some critical comments and suggestions for future research can be given. First, careful reflection is needed on the candidate surrogate endpoints that are considered in the analyses. Indeed, only endpoints for which there is substantial evidence of a causal relationship with the treatment effects on T (in terms of the temporal, biological, and/or pathological association) should be considered as candidate-surrogates the analysis (Buyse *et al.*, 2022; Ciani *et al.*, 2017). The use of a more exploratory approach where a large number of candidate surrogate endpoints (that have no clear causal association with T) are 'tried out' should be avoided – because such an approach could result in an over-fitted model (due to false positives/Type I errors). Note also that the number of available clinical trials (or alternative clustering units such as treating physician) sets limits to the number of surrogate endpoints that can be jointly considered. For example, when $N = 5$ and a full fixed-effects modelling approach is used, at most 2 surrogates can be considered in the Stage 2 analysis.

On a related note, the R^2_{trial} and R^2_{indiv} estimates cannot decrease when additional surrogates are added to the existing ones (as these metrics of surrogacy are essentially coefficients of determination). The question may rise how large the increase in R^2_{trial} and R^2_{indiv} should be to retain the additional candidate surrogate endpoint(s). This decision can be based on formal statistical tests or informal information criteria (such as the LR-tests and the AIC/BIC that were used in the case study), but more qualitative arguments can also be taken into consideration. For example, adding additional surrogates may increase the burden to the patients and/or the research nurses, the financial cost of conducting the clinical trials, and the probability of having missing values. Experts in the field should carefully balance the costs of considering additional surrogates against their benefits in terms of increased prediction accuracy (i.e., the increase in the R^2_{trial} and R^2_{indiv} metrics). Assam *et al.* (2010) formalized such an approach in a longitudinal meta-analytic surrogate evaluation context. In particular, these authors specified an objective function that con-

tained (i) a prediction accuracy component and (ii) a cost component (i.e., the financial cost per additional repeated measurement of the surrogate). The objective function was subsequently maximized to determine the optimal number of measurements, where the importance of the prediction accuracy and the cost components can be gauged through the use of weights. Similar ideas could be used in the multiple surrogacy setting that is considered in the current paper, but this goes beyond the scope of the present work.

Second, the question may arise as to how high R^2_{trial} and R^2_{indiv} should be to conclude that $\boldsymbol{S}$ is "valid". It is however difficult to justify the use of a one-size-fits-all general cut-off. Indeed, the appropriateness and usefulness of $\boldsymbol{S}$ does not only depend on its R^2_{trial} and R^2_{indiv} values, but also on less formal considerations such as the time that is gained by using $\boldsymbol{S}$ instead of T, the reduction in the burden/pain to the patient when considering $\boldsymbol{S}$ instead of T, and so on. For example, Buyse *et al.* (2000) showed that Progression Free Survival (PFS) is an excellent surrogate for Overall Survival (OS) in ovarian cancer (for the treatments cyclophosphamide and cisplatin versus cyclophosphamide, adriamycin, and cisplatin), with $R^2_{trial} = 0.940$ and $R^2_{indiv} = 0.951$. These authors nonetheless concluded that the real-life usefulness of PFS as a surrogate for OS in ovarian cancer is quite limited because both events occur in close temporal proximity to each other. On the other hand, a surrogate with a more moderate R^2_{trial} and R^2_{indiv} can be very useful in practice when it can be measured substantially earlier than the true endpoint. For example, consider a clinical trial in early Alzheimer's disease where T = change in cognition 5 years after the start of treatment and $\boldsymbol{S}$ = tau pathology in different brain areas 1 year after the start of treatment. Even if the R^2_{trial} and R^2_{indiv} metrics would be of a moderate magnitude (say, around 0.60), $\boldsymbol{S}$ could still be useful in practice because it would substantially reduce the time that is needed to conduct future clinical trials. So, instead of specifying a general threshold value for R^2_{trial} and R^2_{indiv} that should always be exceeded to conclude that a (multiple) surrogate endpoint is "valid", it seems to make more sense that such threshold values are determined on a case-by-case basis where medical experts, regulatory agencies, and statistical experts provide input.

Third, an important difference between the single and multiple surrogate endpoint evaluation frameworks is that model-building considerations become more prominent in

the latter approach. Indeed, in the single surrogate setting only one S is considered and thus no model-building is required (i.e., it is essentially assumed that the (treatment effects on) S and T are linearly related – even though non-linear relations can be considered as well in the single surrogate setting, see e.g., Assam *et al.*, 2007). In the multiple surrogate endpoint evaluation framework, several model-building strategies can be considered. A forward selection approach was used in the case study analysis, but alternatively e.g., a backward selection procedure or least absolute shrinkage and selection operator (LASSO) regression could have been considered to conduct the model-building exercise. The use of the latter method is illustrated in the online Appendix for the case study analysis.

On a related note, in the fixed-effects modelling approach the R^2_{trial} value is obtained by fitting model (7). It is thus assumed that the trial-specific treatment effects on T can be predicted based on a linear combination of the main effects of the trial-specific intercepts and treatment effects on $\boldsymbol{S}$ (the same assumption is essentially made in the mixed-effects modelling approach where R^2_{trial} is estimated based on the $\mathbf{D}$ matrix). Depending on the setting at hand, more complex models might be considered. For example, in oncology the effect of a treatment often depends on the treatment's efficacy and its toxicity. In such a setting, it might be sensible to include an interaction term between the trial-specific treatment effects on the efficacy and toxicity surrogate endpoints in the Stage 2 model. An example is provided in the online Appendix.

Fourth, the case study is somewhat unusual in the sense that $\boldsymbol{S}$ and T are based on the same PANSS items (i.e., the PANSS Total score is the sum of the items of the 5 subscales plus 5 additional items that are not part of these subscales), and are measured at the same time. In most surrogacy settings, $\boldsymbol{S}$ and T are different endpoints (e.g., $\boldsymbol{S}$ = tau pathology and T = cognition) and $\boldsymbol{S}$ is measured before T. The PANSS case study is thus a bit atypical, but it was nonetheless used to illustrate the methodology in the current paper because (i) the data are in the open domain, and (ii) this case study has already been analysed in previous publications using different surrogate endpoint evaluation methods. This has the advantages (i) that the results can be easily shared/published, and (ii) that the current results can be compared with the results of previous analyses. For example, Flórez *et al.* (2022) also used the PANSS case study to study multiple surrogacy using

causal-inference methodology in the single-trial setting (one of the 5 clinical trials was used in their analysis). These authors used the so-called Individual Causal Association (ICA) to quantify the strength of the association between the individual causal treatment effects on $\boldsymbol{S}$ and T (Van der Elst *et al.*, 2019). The results of Flórez *et al.* (2022) are fully in line with the current results. To illustrate this, Table 4 shows the medians of the distributions of the ICA for different combinations of the PANSS subscales that were reported in Flórez *et al.* (2022). Note that *distributions* of ICA are obtained (rather than point estimates), because these authors used a sensitivity analysis to deal with the non-identifiability issues that are encountered in the causal-inference framework. The table also shows the R^2_{trial} and R^2_{indiv} estimates that were obtained using the full multivariate weighted fixed-effects modelling approach. As can be seen, the median ICA values that were reported in Flórez *et al.* (2022) are close to the R^2_{trial} and R^2_{indiv} estimates. Both analyses thus lead to the same conclusions with respect to the appropriateness of $\boldsymbol{S}$, despite the fundamentally different statistical frameworks that are used in both approaches and their different assumptions (sensitivity analysis).

<Insert Table 4 about here>

Fifth, different versions of the PANSS have been described in the literature (for an overview, see Lindenmayer, 2017). The clinical trials that were considered in the case study all used the same version of the PANSS (i.e., the original 30 item version that was proposed by Lindenmayer *et al.*, 1995). In a situation where different PANSS versions are used, potential version effects should be taken into account in the analyses (as otherwise e.g., the trial-specific estimated treatment effects would no longer be comparable across clinical trials). For example, conversion equations can be used to transform the PANSS (subscale) scores across different versions (Grot *et al.*, 2021). For psychometric scales for which such conversion equations are not available in the literature, Item Response Theory-based test equating methods can be used to make the different test versions more comparable. Indeed, a key property of Item Response Theory is that the latent patient traits (here: symptom severity levels of schizophrenia) are item – and thus test version – *in*dependent (for details, see Embretson and Reise, 2000; Van der Elst *et al.*, 2013).

Notice that issues with the lack of comparability of different measurement systems are

not restricted to psychometric scales such as the PANSS, but can also be encountered in various other settings. For example, the assays and laboratory protocols that are used to measure COVID-19 antibodies or tau pathology are poorly standardized (Goldblatt *et al.*, 2022; Maass *et al.*, 2017). For such endpoints, differences in the measurement systems also have to be taken into account (by e.g., using conversion equations).

Finally, the methodology that was proposed in the present paper only deals with the setting in which all endpoints are continuous normally distributed variables. The generalisation of the methodology to other types of endpoints – i.e., any mix of binary, ordinal, continuous normally distributed and time-to-event endpoints – is not straightforward. In general, trial-level surrogacy can still be estimated with some small modifications of the methodology. For example, suppose that all endpoints are binary. In such a scenario, the univariate fixed-effects approach (see Section 3) can be adapted by fitting logistic regression models for all endpoints (in Stage 1), after which the obtained trial-specific treatment effects on T are regressed on the trial-specific treatment effects on $\boldsymbol{S}$ (in Stage 2) to estimate R^2_{trial}. The estimation of individual-level surrogacy however becomes more challenging in non-normal settings. Indeed, for continuous normally distributed endpoints, it is natural to quantify R^2_{indiv} based on the association between the residuals. Such a metric is however no longer meaningful for non-normal endpoints. Depending on the particular combination of the endpoints at hand, several approaches can be used to quantify R^2_{indiv}. For example, in the single surrogate setting where both S and T are binary, R^2_{indiv} has been defined as the squared correlation between two latent normally distributed variables that are assumed to underlie the binary outcomes, or alternatively as the global odds ratio between the binary endpoints as estimated using a bivariate Plackett-Dale model (Renard *et al.*, 2002).

Different types of endpoints thus require different definitions of surrogacy in the meta-analytic framework (in particular for the individual-level metrics), which do not necessarily have the same interpretation or even lead to the same conclusion with respect to the appropriateness of S (Alonso *et al.*, 2016). Fortunately, Alonso and Molenberghs (2007) proposed an information-theoretic approach that allows for estimating individual-level surrogacy (as well as trial-level surrogacy) for different types of endpoints in a unified way. This method is highly flexible, and it can e.g., also be used in a setting where S and/or T are longitudinal

measurements (Alonso *et al.*, 2003). The information-theoretic approach can also be useful in the setting where multiple non-normally distributed endpoints are considered, but this is beyond the scope of the current paper.

Supplementary Materials

The methodology that is proposed in this manuscript is implemented in the R package *Surrogate* (available for download at `https://CRAN.R-project.org/package=Surrogate`). An online Appendix that details the analysis of the case study using the R package *Surrogate* is also available.

Conflict of interest

Wim Van der Elst, Helena Geys and Lewin Eisele are employees of and hold shares in Johnson & Johnson.

Acknowledgements

Fenny Ong gratefully acknowledges the support from the Special Research Fund (BOF) of Hasselt University (BOF-number: BOF2OCPO3) and GlaxoSmithKline Biologicals for this study. Florian Stijven gratefully acknowledges the support from Baekeland Mandaat (HBC.2022.0145) and Johnson & Johnson Innovative Medicine.

References

Alonso, A., Geys, H., Molenberghs, G., and Kenward, M. G. (2003), "Validation of surrogate markers in multiple randomized clinical trials with repeated measures," *Biometrical Journal*, 45, 931-945.

Alonso, A., Geys, H., Molenberghs, G., Kenward, M. G., and Vangeneugden, T. (2004), "Validation of surrogate markers in multiple randomized clinical trials with repeated measurements: canonical correlation approach," *Biometrics*, 60, 845-53.

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

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

Alonso, A. A., Bigirumurame, T., Burzykowski, T., Buyse, M., Molenberghs, G., Muchene, L., Perualila, N. J., Shkedy, Z., and Van der Elst, W. (2016), *Applied surrogate endpoint evaluation methods with SAS and R*, New York: CRC Press.

American Psychiatric Association (2000), *Diagnostic and Statistical Manual of Mental Diseases*, Washington, DC: American Psychiatric Association.

Assam, N. P., Tilahun, E. A., Alonso, A. A., and Molenberghs, G. (2007), "Information-theory based surrogate marker evaluation from several randomized clinical trials with continuous true and binary surrogate endpoints," *Clinical trials*, 4, 587–597.

Assam, N. P., Tilahun, E. A., Alonso, A. A., and Molenberghs, G. (2010), "Using earlier measures in a longitudinal sequence as a potential surrogate for a later one," *Computational Statistics & Data Analysis*, 54, 1342–1354.

Bejanin, A., Schonhaut, D. R., La Joie, R., Kramer, J. H., Baker, S. L., Sosa, N., Ayakta, N., Cantwell, A., Janabi, M., Lauriola, M., O'Neil, J. P., Gorno-Tempini, M. L., Miller, Z. A., Rosen, H. J., Miller, B. L., Jagust, W. J., and Rabinovici, G. D. (2017), "Tau

pathology and neurodegeneration contribute to cognitive impairment in Alzheimer's disease," *Brain*, 140, 3286–3300.

Blin, O., Azorin, J. M., and Bouhours, P. (1996), "Antipsychotic and anxiolytic properties of risperidone, haloperidol and methotrimeprazine in schizophrenic patients," *Journal of Clinical Psychopharmacology*, 16, 38–44.

Burzykowski, T., Molenberghs, G., and Buyse, M. (2005), *The Evaluation of Surrogate Endpoints*, New York: Springer-Verlag.

Buyse, M., and Molenberghs, G. (1998), "The validation of surrogate endpoints in randomized experiments," *Biometrics*, 54, 1014–1029.

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

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

Buyse, M., Saad, E. D., Burzykowski, T., Regan, M. M., and Sweeney, C. S. (2022), "Surrogacy beyond prognosis: the importance of 'trial-level' surrogacy," *The Oncologist*, 27, 266–271.

Ciani, O., Buyse, M., Drummond, M., Rasi, G., Saad, E. D., and Taylor, R. S. (2017), "Time to review the role of surrogate end points in health policy: state of the art and the way forward", *Value in Health*, 20, 487–495.

Chouinard, G., Jones, B., and Remington, G. (1993), "A Canadian multicenter placebo-controlled study of fixed doses of risperidone and haloperidol in the treatment of chronic schizophrenic patients, " *Journal of Clinical Psychopharmacology*, 13, 25–40.

Cortiñas Abrahantes, J., Molenberghs, G., Burzykowski, T., Shkedy, Z., and Renard D. (2004). Choice of units of analysis and modeling strategies in multilevel hierarchical models. *Computational Statistics and Data Analysis, 47,* 537–563.

Cortiñas, A. J., Shkedy, Z., and Molenberghs, G. (2008), "Alternative methods to evaluate trial level surrogacy," *Clinical trials*, 5, 194–208.

Embretson, S. E., and Reise, S. P. (2000), *Item response theory for psychologists*, Mahwah, NJ: Lawrence Erlbaum.

Flórez, A. J., Molenberghs, G., Van der Elst, W., Alonso, A. A. (2022), "An efficient algorithm to assess multivariate surrogate endpoints in a causal inference framework," *Computational Statistics & Data Analysis*, 172, 1–12.

Freedman L. S., Graubard, B. I., and Schatzkin, A. (1992), "Statistical validation of intermediate endpoints for chronic diseases," *Statistics in Medicine*, 11, 167–178.

Goldblatt, D., Fiore-Gartland, A., Johnson, M., Hunt, A., Bengt, C., Zavadska, D., Snipe, H. D., Brown, J. S., Workman, L., Zar, H. J., Montefiori, D., Shen, X., Dull, P., Plotkin, S., Siber, G., and Ambrosino, D. (2022), "Towards a population-based threshold of protection for COVID-19 vaccines," *Vaccine*, 40, 306–315.

Grot S., Giguére C., Smine S., Mongeau-Pérusse V., Diem Nguyen D., Preda A., Potvin S., van Erp T. G. M., Orban P. (2021), "Converting scores between the PANSS and SAPS/SANS beyond the positive/negative dichotomy," *Psychiatry Research*, 305.

Hoyberg, O. J., Fensbo, C., Remvig, J., Lingjaerde, O. K., Slotei-Nielsen, M., and Salvesen, I. (1993), "Risperidone versus perphenazine in the treatment of chronic schizophrenic patients with acute exacerbations," *Acta Psychiatrica Scandinavica*, 13, 395–402.

Huttunen, M. O., Piepponen, T., Rantanen, H., Larmo, I., Nyholm, R., and Raitasuo, V. (1995), "Risperidone versus zuclopenthixol in the treatment of acute schizophrenic episodes: a double-blind parallel-group trial," *Acta Psychiatrica Scandinavica*, 91, 271–277.

Johnson, R. A., and Wichern, D. W. (2007), *Applied Multivariate Statistical Analysis*, New Jersey: Pearson Prentice-Hall.

Lee, Y. S. (1971), Tables of the upper percentage points of the multiple correlation. *Biometrika*, 59, 175–189.

Lindenmayer, J. P., Bernstein-Hyman, R., Grochowski, S., and Bark, N. (1995), "Psychopathology of schizophrenia: initial validation of a 5-factor model," *Psychopathology*, 28, 22–31.

Lindenmayer, J.P. (2017), "Are Shorter Versions of the Positive and Negative Syndrome Scale (PANSS) Doable? A Critical Review," *Innov Clin Neurosci*, 14, 11–12.

Lindstrom, M. J., and Bates, D. M. (1988), "Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data," *Journal of the American Statistical Association*, 83, 1014–1022.

Maass, A., Landau, S., Baker, S. L., Horng, A., Lockhart, S. N., La Joie, R., Rabinovici, G. D., Jagust, W. J. (2017), "Comparison of multiple tau-PET measures as biomarkers in aging and Alzheimer's disease", *NeuroImage*, 157, 448–463.

Ong, F., Wang, J., Van der Elst, W., Verbeke, G., Molenberghs, G., and Alonso, A. A. (2022), "Implementing the meta-analytic approach for the evaluation of surrogate endpoints in SAS and R: a word of caution," *Journal of Biopharmaceutical Statistics*, 32, 705–716.

Parast, L., Cai, T., and Tian, L. (2021), "Evaluating multiple surrogate markers with censored data," *Biometrics*, 77, 1315–1327.

Peuskens, J. and the Risperidone Study Group (1995), "Risperidone in the treatment of patients with chronic schizophrenia: a multi-national multi-centre, double blind, parallel groups study versus haloperidol," *British Journal of Psychiatry*, 166, 712–726.

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

Renard, D., Geys, H., Molenberghs, G., Burzykowski, T., and Buyse, M. (2002), "Validation of surrogate endpoints in multiple randomized clinical trials with discrete outcomes," *Biometrical Journal*, 44, 921–935.

Singh, M., and Kay, S. (1975), "A comparative study of haloperidol and chlorpromazine in terms of clinical effects and therapeutic reversal with benztropine in schizophrenia. Theoretical implications for potency differences amongst neuroleptics," *Psychopharmacologia*, 43, 103–113.

Tibaldi, F. S., Cortiñas Abrahantes, J., Molenberghs, G., Renard, D., Burzykowski, T., Buyse, M., Parmar, M., Stijnen, T., and Wolfinger, R. (2003), "Simplified hierarchical linear models for the evaluation of surrogate endpoints," *Journal of Statistical Computation and Simulation*, 73, 643–658.

Van der Elst, W., Ouwehand, C., van Rijn, P., Lee, N., Van Boxtel, M. P. J., and Jolles, J. (2013), "The shortened Raven Standard Progressive Matrices: Item Response Theory-based psychometric analyses and normative data. Assessment," *Assessment*, 20, 48–59.

Van der Elst, W., Hermans, L., Verbeke, G., Kenward, M. G., Nassiri, V., and Molenberghs, G. (2016), "Unbalanced cluster sizes and rates of convergence in mixed-effects models for clustered data," *Journal of Statistical Computation and Simulation*, 86, 2123–2139.

Van der Elst, W., Alonso, A. A., Geys, H., Meyvisch, P., Bijnens, L., Sengupta, R., and Molenberghs, G. (2019), "Univariate versus multivariate surrogate endpoints in the single-trial setting," *Statistics in Biopharmaceutical Research*, 3, 301–310.

Verbeke, G., and Molenberghs, G. (2000), *Linear Mixed Models for Longitudinal Data*, New York: Springer-Verlag.

Xu, J., and Zeger, L. Z. (2001), "The evaluation of multiple surrogate endpoints," *Biometrics*, 57, 81–87.

Tables

Table 1: Case study. Pearson correlations between the PANSS subscale scores and the PANSS Total score in the active control (see panel [a]) and experimental treatment (see panel [b]) arms.

[a] Active control

	Neg	Exc	Cog	Pos	Dep	Total
Neg	1.000	0.291	0.569	0.301	0.379	0.722
Exc	0.291	1.000	0.509	0.678	0.561	0.758
Cog	0.569	0.509	1.000	0.502	0.405	0.805
Pos	0.301	0.678	0.502	1.000	0.584	0.772
Dep	0.379	0.561	0.405	0.584	1.000	0.723
Total	0.722	0.758	0.805	0.772	0.723	1.000

[b] Experimental treatment

	Neg	Exc	Cog	Pos	Dep	Total
Neg	1.000	0.338	0.639	0.336	0.376	0.764
Exc	0.338	1.000	0.501	0.653	0.609	0.754
Cog	0.639	0.501	1.000	0.542	0.339	0.812
Pos	0.336	0.653	0.542	1.000	0.552	0.775
Dep	0.376	0.609	0.339	0.552	1.000	0.693
Total	0.764	0.754	0.812	0.775	0.693	1.000

Note. Neg = Negative symptoms, Pos = Positive symptoms, Cog = Cognitive symptoms, Exc = Excitement, Dep = Depression, and Total = PANSS Total score.

Table 2: Results of the case study analysis. Estimated R^2_{trial} and R^2_{indiv} (and their 95% CI) that are obtained when 1, 2,..., and 5 surrogate endpoints (i.e., the 5 PANSS subscales) are used (with T = PANSS Total score) based on a full multivariate weighted fixed-effects modelling approach. The table shows the results for all single surrogates and for the best nested combination of 2 to 5 surrogate endpoints.

Surrogates considered	R^2_{trial} (and 95% CI)	R^2_{indiv} (and 95% CI)	AIC	BIC	LR-test
Neg	0.600 [0.475; 0.695]	0.567 [0.537; 0.595]	788.3	799.7	
Pos	0.644 [0.527; 0.731]	0.599 [0.571; 0.626]	773.8	785.1	
Exc	0.594 [0.468; 0.690]	0.549 [0.518; 0.578]	790.3	801.6	
Cog	0.658 [0.544; 0.742]	0.652 [0.626; 0.676]	768.6	780.0	
Dep	0.450 [0.308; 0.567]	0.487 [0.454; 0.518]	828.5	839.9	
Cog and Dep	0.866 [0.808; 0.901]	0.831 [0.816; 0.844]	654.8	671.9	Compare with S = Cog: $\chi^2 = 117.8$ (DF = 2) with p-value < 0.001
Cog, Dep and Neg	0.923 [0.887; 0.943]	0.894 [0.884; 0.902]	588.2	610.9	Compare with $\boldsymbol{S}$ = Cog and Dep: $\chi^2 = 70.6$ (DF = 2) with p-value < 0.001
Cog, Dep, Neg and Pos	0.975 [0.962; 0.981]	0.969 [0.966; 0.971]	449.7	478.1	Compare with $\boldsymbol{S}$ = Cog, Dep and Neg: $\chi^2 = 142.5$ (DF = 2) with p-value < 0.001
Cog, Dep, Neg, Pos and Exc	0.989 [0.983; 0.992]	0.990 [0.989; 0.991]	350.7	384.7	Compare with $\boldsymbol{S}$ = Cog, Dep, Neg and Pos: $\chi^2 = 103.1$ (DF = 2) with p-value < 0.001

Note. Neg = Negative symptoms, Pos = Positive symptoms, Cog = Cognitive symptoms, Exc = Excitement and Dep = Depression.

Table 3: Results of the simulation study. Bias, efficiency, and coverage for R^2_{trial} and R^2_{indiv} in the full multivariate mixed- and fixed-effects modelling approaches.

Model fitted		R^2_{trial}			R^2_{indiv}		
	N	Bias	Efficiency	Coverage	Bias	Efficiency	Coverage
Full multivariate mixed-effects model	10	0.220	0.151	92.3%	0.008	0.055	92.3%
	20	0.139	0.140	93.9%	0.004	0.036	94.8%
	50	0.053	0.094	94.9%	< 0.001	0.022	95.4%
Full multivariate fixed-effects model	10	0.315	0.143	93.9%	0.008	0.051	93.7%
	20	0.135	0.137	94.1%	0.003	0.036	94.9%
	50	0.052	0.092	95.3%	< 0.001	0.022	95.3%

Table 4: Comparison of the results of the surrogacy analyses of Flórez *et al.* (2022) and the current analyses.

Surrogates considered	Results Flórez *et al.* (2022) ICA	Results current paper R^2_{trial}	R^2_{indiv}
Pos	0.641	0.644	0.599
Pos, Cog	0.821	0.815	0.825
Neg, Pos, Cog	0.919	0.932	0.925
Neg, Pos, Cog, Dep	0.960	0.975	0.969
Neg, Pos, Exc, Cog, Dep	0.992	0.989	0.990

Note. Neg = Negative symptoms, Pos = Positive symptoms, Cog = Cognitive symptoms, Exc = Excitement and Dep = Depression.

Figures

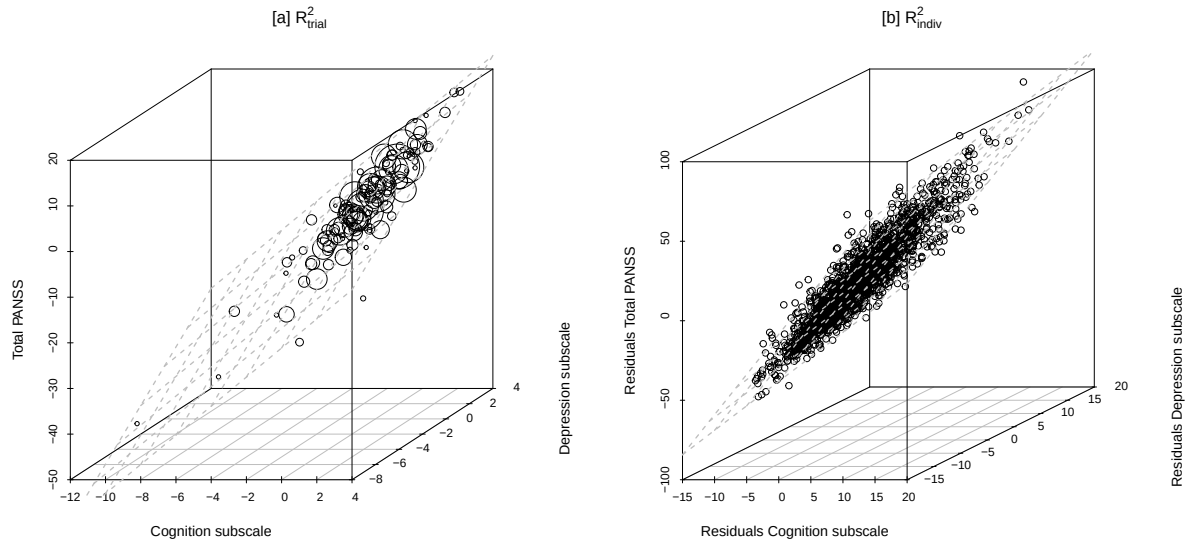

Figure 1: Visual illustration of the R^2_{trial} and R^2_{indiv} results for the analysis with a multiple surrogate $\boldsymbol{S}$ that consists of the Cognition and the Depression subscales. Panel [a] shows the physician-specific treatment effects on the Cognition subscale, the Depression subscale, and the PANSS Total score (see the black circles; the circumference of the circles is proportional to the number of patients that are treated by a physician), supplemented with a fitted regression plane that corresponds to a reduced weighted Stage 2 model (as the full model cannot be shown in 3 dimensions; see the grey dashed lines). Panel [b] shows the residuals of the Cognition subscale, the Depression subscale, and the PANSS Total score, supplemented by a fitted regression plane (see the black circles and grey dashed lines, respectively).

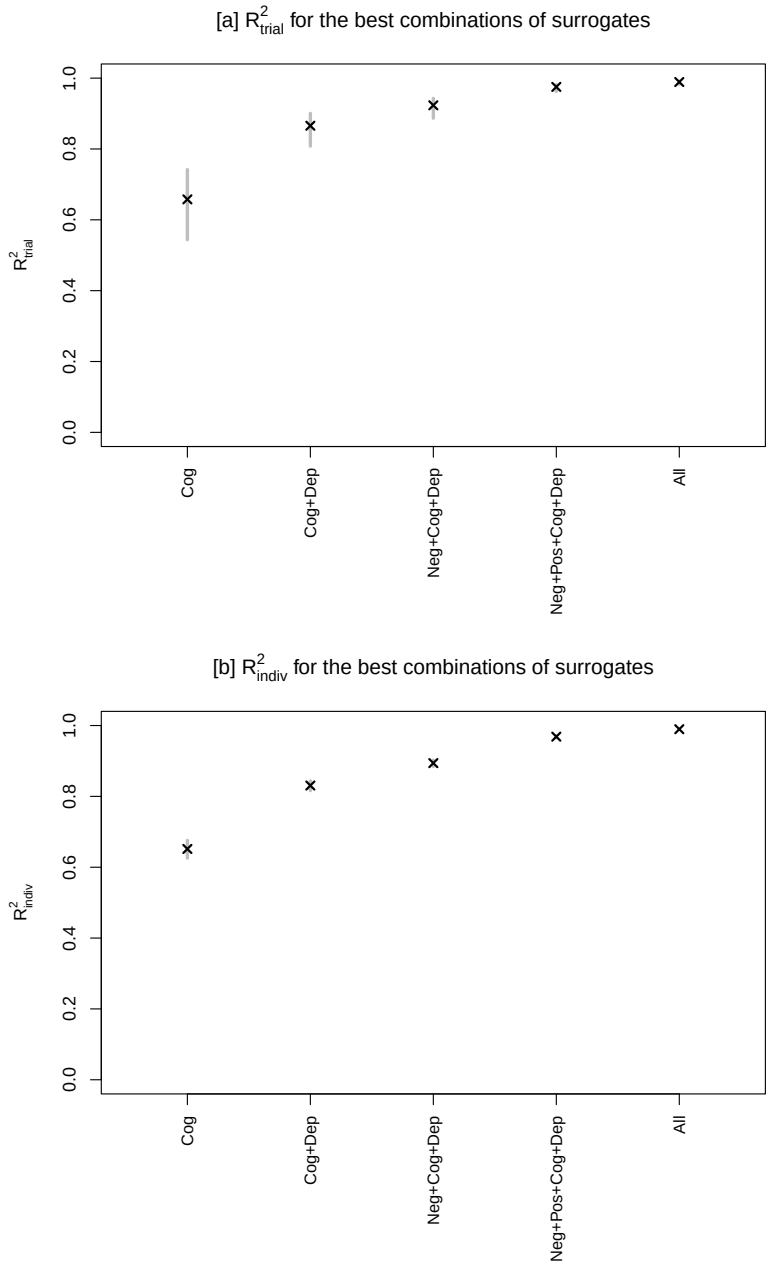

Figure 2: Results of the case study analysis. Estimated R^2_{trial} (panel [a]) and R^2_{indiv} (panel [b]) for the best combinations of 1, 2, ..., and 5 surrogates. The black crosses and grey lines correspond to the point estimates and their 95% CIs, respectively.

Note. Neg = Negative symptoms, Pos = Positive symptoms, Cog = Cognitive symptoms, and Dep = Depression.