Investigating the criterion validity of psychiatric symptom scales using surrogate marker validation methodology

ALONSO ABAD, Ariel; GEYS, Helena; MOLENBERGHS, Geert & VANGENEUGDEN, Tony (2002) Investigating the criterion validity of psychiatric symptom scales using surrogate marker validation methodology. In: Journal of Biopharmaceutical Statistics, 12(2). p. 161-179.

DOI: 10.1081/BIP-120015741
Handle: http://hdl.handle.net/1942/409

Investigating the Criterion Validity of Psychiatric Symptom Scales using Surrogate Marker Validation Methodology

Ariel Alonso

Limburgs Universitair Centrum, Center for Statistics, Biostatistics, Universitaire Campus, B-3590 Diepenbeek, Belgium

Tel: ++32/(0)11/26.82.82

Fax: ++32/(0)11/26.82.99

Email: ariel.alonso@luc.ac.be

Helena Geys

Limburgs Universitair Centrum, Center for Statistics, Biostatistics, Universitaire Campus, B-3590 Diepenbeek

Belgium

Tel: ++32/(0)11/26.82.35

Fax: ++32/(0)11/26.82.99

Email: helena.geys@luc.ac.be

Geert Molenberghs

Limburgs Universitair Centrum, Center for Statistics, Biostatistics, Universitaire Campus, B-3590 Diepenbeek

Belgium

Tel: ++32/(0)11/26.82.38

Fax: ++32/(0)11/26.82.99

Email: geert.molenberghs@luc.ac.be

Tony Vangeneugden

Janssen Research Foundation, Turnhoutseweg 30, B-2340 Beerse,

Belgium

Tel: ++32/(0)14/60.35.95

Fax: ++32/(0)14/60.54.83

Email: tvangene@janbe.jnj.com

Abstract

This work investigates whether techniques that are generally used for the validation of surrogate markers in clinical trials, can be applied in the validation of psychiatric health measurements (often scales) and more generally to investigate relationships between treatment effects on different measurements. When psychiatric health measurements are either developed or used in a new population, reliability and validity must be investigated. Reliability, more specifically internal consistency, test-retest reliability and inter-rater reliability, is focused on the reproducibility of the measurement. Validity on the other hand, is defined as the degree to which the scale measures what it purports to measure. This can be performed through the analysis of content, construct and criterion validity. We argue that recent methodology, in particular developed to study surrogate endpoints, can be used to examine criterion validity, concurrent validity and predictive validity. In concurrent validity, we correlate the measurement with a criterion measure, both of which are given at the same time. In predictive validity, the criterion will not be available to some point in time in the future. The surrogate methods were applied on pooled data from 5 trials in schizophrenia.

Keywords

Criterion Validity; Concurrent Validity; Meta-analysis; Psychiatry; Predictive Validity; Surrogate Endpoint.

1 INTRODUCTION

In this paper we illustrate how recently proposed criteria for the validation of surrogate markers in clinical trials can be easily adapted and used to assess the so-called *criterion validity* of psychiatric symptom scales. This concept will be described further in this paper.

One feature of psychiatric health sciences literature devoted to measuring subjective states is the daunting area of available scales (Streiner (1) and Norman). The development of scales to assess subjective

attributes is not easy and subject to many controversial debates. One particular drawback of course lies in the fact that the filling-in of a scale may vary from one person to another. Because of the subjective nature of many of these scales, one may encounter scales that are not adequate to assess a particular concept. Therefore, whenever a mental health measurement scale is developed or translated or used in a new population, its psychometric properties have to be assessed. Two important properties are *reliability* and *validity*.

Reliability consists in determining the extent to which the measurement is free from random error. This can be performed through analysing *internal consistency* and *reproducibility* of the questionnaire. Internal consistency is the extent to which individual items are consistent with each other and reflect a single underlying construct. Essentially, internal consistency represents the average of the correlations among all the items in the instrument. Several measures that are often used to provide proof of internal consistency are: Cronbach's alpha coefficient (Cronbach(2)), Kuder-Richardson (Kuder(3) and Richardson) and factorial analyses. Intra-observer or test-retest reliability is the degree to which a measure yields stable scores at different points in time for patients who are assumed not to have changed clinical status on the domains being assessed. The calculation of intraclass correlation coefficients (Fleiss(4) and Cohen, Deyo(5), Dierh and Patrick) is one of the most commonly used methods. For interviewer-administered questionnaires, the inter-observer reliability is the degree to which a measurement yields stable scores when administered by different interviewers, rating the same patients. The calculation of interclass correlation coefficients is also one of the most commonly used methods.

The validity of a questionnaire is defined as the degree to which the questionnaire measures what it purports to measure. This can be performed through the analysis of *content*, *construct* and *criterion validity*. Content validity can be defined as the extent to which the instrument assesses all the relevant or important content or domains. Also the term *face validity* is used to indicate whether the instrument appears to be assessing the desired qualities at face. This form of validity consists of a judgement by experts in the field. Construct validity refers to a wide range of approaches which are used when what we are trying to measure is a "hypothetical construct" (e.g., anxiety, irritable bowel syndrome, ...) rather than something that can be readily observed. The most commonly used methods to explore construct validity are: extreme groups (apply instrument for example to cases and non-cases), convergent and discriminant validity testing (correlate with other measures of this construct and not correlate with disimilar or unrelated constructs) and multitrait-multimethod matrix (Campbell(6) and Fisk). Criterion validity can be divided into two

types: *concurrent validity* and *predictive validity*. With concurrent validity we correlate the measurement with a criterion measure (gold standard), both of which are given at the same time. In predictive validity, the criterion will not be available until some time in the future. The most commonly used method to assess the validity is by calculation of the Pearson correlation coefficient.

The idea of the present work is to provide a new way of investigating criterion validity of psychiatric symptom scales, using the criteria applied in surrogate marker validation for clinical trials. Surrogate endpoints are loosely referred to as endpoints that can be used instead of other endpoints in the evaluation of experimental treatments or other interventions. The validation of surrogate endpoints is a controversial issue (Boissel(7), Collet, Moleur and Haugh; Fleming(8) and DeMets; De Gruttola(9), Fleming, Lin and Coombs) and should be rigorously established. In a landmark paper, Prentice(10) proposed a formal definition of surrogate endpoints and outlined how potential surrogate endpoints could be validated. Much debate ensued, for the criteria set out by Prentice are too stringent (Fleming(11), Prentice, Pepe and Glidden) and neither necessary nor sufficient for his definition to be fulfilled, except in the special case of binary outcomes (Buyse(12) and Molenberghs). In addition, Freedman(13), Graubard and Schatzkin showed that these criteria were not straightforward to verify through statistical hypothesis tests. They introduced the *proportion explained* (PE) to quantify how much of the treatment effect is captured by the surrogate endpoint. The latter proposal is itself surrounded with difficulties, the most dramatic one being that it is not confined to the unit interval (Molenberghs(14), Buyse, Burzykowski, Renard and Geys). Buyse(12) and Molenberghs proposed to replace PE by two new measures to assess the quality of a surrogate. The first one, termed *relative effect* (RE) is the (population-averaged) effect of the treatment on the true endpoint relative to that on the surrogate endpoint. The second one is the *adjusted association* between both endpoints, an individual measure of agreement between both endpoints after accounting for the effect of treatment. Technically, a joint model for both endpoints is required. In turn, a drawback of the RE is that, when calculated from a single trial, its use depends on strong unverifiable assumptions, the main one being that it should be constant across a class of trials. A way out of this problem is the combination of information from several groups of patients (multi-center trials or meta-analyses). Such an approach was suggested by Albert(15) *et al.*, and was implemented by Daniels(16) and Hughes, Buyse(17), Molenberghs, Burzykowski, Renard and Geys and Gail(18), Pfeiffer, Van Houwelingen and Carroll. Buyse(17) *et al.* show that the individual-level association between the surrogate and final endpoints carries over naturally to this setting. The notion of relative effect, on the other hand, can be extended to a trial-level measure of

association between the effects of treatment on both endpoints. Their approach suggests a new definition of validity in terms of the quality of both trial-level and individual-level associations between the surrogate and true endpoints. The quality of a surrogate at the trial level is assessed by means of a coefficient of determination R^2_{trial}. At the individual level the squared correlation R^2_{indiv} between the surrogate and true endpoint, after adjustment for both the trial effects and the treatment effects is used. A surrogate will be said to be valid when it is both trial-level valid ($R^2_{trial} \approx 1$) and individual-level valid ($R^2_{indiv} \approx 1$). From a modelling perspective, a two-stage hierarchical model is required. This can be fitted using a variety of methods, such as linear mixed-effects methodology (Verbeke(19) and Molenberghs), a two-stage approach, or pseudo-likelihood (Geys(20)). Several methods have been proposed for applications in different settings. For example, Molenberghs(21), Geys and Buyse developed a pseudo-likelihood approach for the validation of a surrogate in a randomized trial when the surrogate and the true outcome are of mixed data types. Renard(22), Geys, Molenberghs, Burzykowski and Buyse extended the meta-analytic setting for two normally distributed outcomes to the case of two binary outcomes using a pseudo-likelihood approach for parameter estimation. Burzykowski(23), Molenberghs, Buyse, Geys and Renard extended the meta-analytic settings for two normally distributed endpoints to the common situation of failure-time endpoints, using bivariate survival modelling. Here, we will show how the meta-analytic approach of Buyse(17) *et al.* can be used to investigate the concurrent validity of two psychiatric rating scales. In cases where a gold standard scale can be assigned, we can almost directly apply their methodology for the validation of surrogate markers with the standard scale playing the role of true endpoint. In many psychiatric studies however, a more "symmetric" situation is encountered where different scales are in conjunction without knowing their relationships. In that case one will need to "symmetrize" the validation techniques. While our data setting does not allow us to investigate the predictive validity, the methods proposed here could be applied to "validate" one scale versus another in that sense as well using clinical trial data.

Section 2 introduces motivating studies on meta-analyses of clinical trials comparing antipsychotic agents for the treatment of chronic schizophrenia. Section 3 gives a brief overview of different validation criteria that exist to validate surrogate endpoints in randomized clinical trials and indicates how these should be adapted to investigate the criterion validity of psychiatric measurement scales. However, it will be pointed out how some of these approaches are surrounded with severe drawbacks and, as a result, may best be avoided. In Section 4 we apply the different methods of Section 3 on the data, described in Section 2. We will show how some of these methods can usefully be applied to investigate the criterion

validity of two rating scales, while others are thus surrounded with difficulties that they may lead to misleading or inconclusive results. The multi-trial approach of Buyse(17) *et al.* will turn out to be really superior. Finally, Section 5 contains some concluding remarks.

2 MOTIVATING STUDIES

2.1 A Meta-analysis of Trials in Schizophrenic Subjects

In this section we introduce individual patient data from a meta-analysis of five double-blind randomized clinical trials, comparing the effects of risperidone to conventional antipsychotic agents for the treatment of chronic schizophrenia. Schizophrenia has long been recognised as a heterogeneous disorder with patients suffering from both "negative" and "positive" symptoms. Negative symptoms are characterized by deficits in cognitive, affective and social functions such as for example poverty of speech, apathy and emotional withdrawal. Positive symptoms entail more florid symptoms such as delusions, hallucinations, and disorganized thinking, which are superimposed on the mental status (Kay(24), Fiszbein and Opler).

Several measures can be considered to assess a patient's global condition. The Clinician's Global Impression (CGI) is generally accepted as a subjective clinical measure of change. Here we will consider the CGI overall change versus baseline. This is a 7–grade scale used by the treating physician to characterize how well a subject has improved since baseline. Other useful and sufficiently sensitive assessment scales are the Positive and Negative Syndrome Scale (PANSS) (Kay(25), Opler and Lindenmayer) and the Brief Psychiatric Rating Scale (BPRS) (Overall(26) and Gorham). The PANSS consists of 30 items that provide an operationalized, drug-sensitive instrument, which is highly useful for both typological and dimensional assessment of schizophrenia (Kay(24), Fiszbein and Opler). The BPRS is a 19-item scale, essentially derived from the PANSS.

Since the package insert in most countries recommend that risperidone is most effective at doses ranging from 4 to 6 mg/day, we included only patients in our analyses that received either these doses of risperidone or an active control (haloperidol, levomepromazine, perphenazine, zuclopenthixol). Depending on the trial, treatment was administered for a duration of 4 to 8 weeks. For example in the international trials (INT-2 by Peuskens(27) and the Risperidone Study Group, INT-3 by Chounard(28), Jones and Remington, and Marder(29) and Meibach, and INT-7 by Hoyberg(30), Fensbo, Remvig, Lingjaerde, Slotei-

Nielsen and Salvesen) patients received treatment for 8 weeks; in the study by Blin(31), Azorin and Bouhours (FRA-3) patients received treatment for 4 weeks, while in the study by Huttunen(32), Piepponen, Rantanen, Larmo, Nyholm and Raitasuo (FIN-1) patients were treated over a period of 6 weeks. In this paper we will restrict our attention to the last observed scores during treatment (endpoint).

Interest is to know to which extent the PANSS and BPRS scales are related with each other and with CGI. We will show that we can use analogous techniques as when validating a surrogate endpoint from meta-analytic data. Our meta-analysis however contains only five trials. This is insufficient to apply the meta-analytic methods of Section 3 (Buyse(17) *et al.*). Fortunately, in all of the trials information is also available on the investigators which treated the patients. Hence, we can also use investigator as the unit of analysis. A total of 138 units are thus available for analysis, with the number of patients per unit ranging from 2 to 30.

2.2 An Equivalence Trial in Schizophrenic Patients

This section describes data from an international equivalence trial (INT-10) on schizophrenic patients, described by Nair(33) and the Risperidone Study Group. The trial includes 206 schizophrenic patients. All patients receive an equal daily amount of risperidone during 8 weeks, but 103 patients are randomized to a one-time daily intake (O.D), while the remaining 103 patients are randomized to receive risperidone twice a day (B.I.D). Like in the previous study, interest lies in determining the extent to which CGI, PANSS and BPRS are related with each other. Since we only had information available on a single trial with one main investigator, we chose to use investigator as the unit of analysis in the multi-trial approach, described in Section 3. A total of 34 units were thus available for analysis with the number of patients per unit ranging from 2 to 15.

3 A BRIEF HISTORY ON VALIDATION CRITERIA

Buyse(12) and Molenberghs have given an overview, with discussion, of common practice for validation of surrogate endpoints. In this section, we summarize their main arguments but in view of assessing the criterion validity of mental health symptom scales.

Let us first introduce some notation. Throughout this chapter we assume that S_1 and S_2 are

random variables that represent two scales for which we want to assess the criterion validity. Traditional approaches investigate the concurrent validity by correlating one measurement scale (S_2) with the other, assumed to be a gold standard (S_1). In many cases an ordinary Pearson's correlation coefficient is used. Here, we propose to assess the criterion validity based on criteria similar to the ones used in surrogate marker validation in randomized clinical trials. In this section, we will give an overview of possible methods that can be applied, however many of them are surrounded with severe difficulties and are thus best avoided. Only the multi-trial approach, described in Section 3.4 will turn out to be really outstanding. While the methods described below could equally well be applied to investigate the predictive validity (where one of the two criteria will not be available until some time in the future), this falls beyond the scope of the data analyses presented in this paper. Further we assume that Z is an indicator variable for treatment. We restrict attention to a binary treatment indicator ($Z = 0$ or 1).

3.1 Prentice's Criteria

Following the ideas of Prentice(10) we assume that criterion validity has been assessed when "the tests of the null hypothesis of no relationship to the treatment groups under comparison are equivalent on either scale":

$$f(S_1|Z) = f(S_1) \Leftrightarrow f(S_2|Z) = f(S_2) \tag{1}$$

where $f(X)$ denotes the probability distribution of a random variable X and $f(X|Z)$ denotes the probability distribution of X conditional on the value of Z. Note that this definition involves the triplet (S_1, S_2, Z), hence concurrent validity between any two scales is assessed only with respect to the effect of some specific treatment Z. Assuming that S_1 can be regarded as the criterion, following 4 validation criteria can be proposed (Prentice(10)):

$$f(S_1|Z) \neq f(S_1), \tag{2}$$

$$f(S_2|Z) \neq f(S_2), \tag{3}$$

$$f(S_1|S_2) \neq f(S_1), \tag{4}$$

$$f(S_1|S_2) = f(S_1|S_2, Z). \tag{5}$$

Criteria (2) and (3) measure departures from the null hypothesis, implicit in (1). Criterion (4) implies that S_2 has prognostic value for the gold standard. Criterion (5) requires S_2 to fully capture the effect of treatment on S_1, that is: there is no effect of treatment on one scale after correction for the other scale.

Of course, this last condition is so restrictive that it rarely holds in practice and it is hard to verify since it would formally require equivalence testing. While in many practical applications one of the symptom scale may be regarded as "the standard", this is not always evident with psychiatric diagnostic tools. In that case we may have to add two extra criteria:

$$f(S_2|S_1) \neq f(S_2), \tag{6}$$

$$f(S_2|S_1) = f(S_2|S_1, Z). \tag{7}$$

Further, in an equivalence trial designed to demonstrate the equivalence of a new treatment with a standard therapy, the first two Prentice criteria are bound not to be fulfilled. Yet, from a clinical perspective there is no reason why the symptom scales used as responses in such a trial cannot be validated. This will be illustrated further in this paper.

3.2 Freedman's Proportion Explained

Freedman(13) *et al.* argued that criterion (5) (and thus also (7)) raises a conceptual difficulty in that it would require the statistical test for treatment effect on one scale to be *non*-significant after adjustment for the other. The non-significance of this test does not prove that the effect of treatment upon the first scale is *fully* captured by the second one. Therefore, they supplemented these criteria with the so-called *proportion explained*, the proportion of the treatment effect on one scale that is explained by the other. Let $PE(S_1, S_2, Z)$ stand for the proportion of the effect of Z on S_1 which can be explained by S_2. An estimate of $PE(S_1, S_2, Z)$ is then as follows:

$$PE(S_1, S_2, Z) = 1 - \frac{\beta}{\beta_{S_2}} \tag{8}$$

where β and β_{S_2} are the estimates of the effect of Z on S_1 without and with adjustment for S_2. Note that this quantity is subject to the same asymetry as criteria (4)-(7). Therefore one might also have to look at $PE(S_2, S_1, Z)$ whenever there is no clear standard among the two considered instruments. Prentice's criterion (5) requires that $\beta_{S_2} = 0$, and thus $PE = 1$ in (8). An instrument for which $PE < 1$ explains only part of the treatment effect on the other instrument. Hence, following the ideas of Freedman(13) *et al.* one could suggest that the criterion validity of two instruments is assessed when the PE is close to unity. In cases where it is not clear which scale can serve as "the standard", both $PE(S_1, S_2, Z)$ and $PE(S_2, S_1, Z)$ should be close to unity. However, this reasoning is not valid. Several conceptual difficulties surrounding the PE have been outlined in the literature (Lin(34), Fleming and De Gruttola, Buyse(12)

and Molenberghs, Flandre(35) and Saidi, Buyse(18) *et al.*, Molenberghs(14) *et al.*), in particular that it is not a proportion: PE can be estimated to be anywhere on the real line, which makes its interpretation problematic.

3.3 Relative Effect and Adjusted Association

Buyse(12) and Molenberghs suggested to replace the PE by two related quantities: the relative effect (RE), which is the ratio of the treatment effects upon the two instruments and the treatment-adjusted association, γ_Z, which is the subject-specific association, adjusted for treatment. Formally, the RE can be written as:

$$RE(S_1, S_2) = \frac{\beta}{\alpha}$$

where β and α are the estimates of the effect of treatment on S_1 and S_2. Note that the RE is anti-symetric in the sense that $RE(S_1, S_2) = 1/RE(S_2, S_1)$, while the adjusted association is fully symetric.

3.4 Multi-trial Approach

Molenberghs(14) *et al.* point to the difficulties accompanying all previous approaches and note that a sensible validation strategy can only be expressed in full in a multi-trial setting. Indeed, serious problems remain in the single trial framework. For instance, when interest lies in predicting the trial-specific treatment effect on S_1 from the treatment effect on S_2, the

$$RE$$

could in principle be used. However, this quantity might not be constant for all trials testing the therapeutic question under consideration. The constancy of RE implies that the relation between α and β is linear through the origin. This assumption may be untenable in practice, and it cannot be verified from a single trial. Therefore Buyse(17) *et al.* adopted an alternative approach based on a meta-analysis of several trials. We will show that this setting is really the most appropriate one for the validation of psychiatric symptom scales.

Let us now present their hierarchical approach. At the first stage, they consider

$$S_{1ij}|Z_{ij} = \mu_{S_{1i}} + \beta_i Z_{ij} + \varepsilon_{S_{1ij}}, \tag{9}$$

$$S_{2ij}|Z_{ij} = \mu_{S_{2i}} + \alpha_i Z_{ij} + \varepsilon_{S_{2ij}}, \tag{10}$$

where α_i and β_i are trial-specific effects of treatment Z on the endpoints in a trial, $\mu_{S_{1i}}$ and $\mu_{S_{2i}}$ are trial-specific intercepts, and $\varepsilon_{S_{1i}}$ and $\varepsilon_{S_{2i}}$ are correlated error terms, assumed to be mean-zero normally distributed with covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_{S_1S_1} & \sigma_{S_1S_2} \\ \sigma_{S_1S_2} & \sigma_{S_2S_2} \end{pmatrix}.$$

Due to the replication at the trial level, they can impose a further model on the trial-specific parameters. At the second stage, they then assume

$$\begin{pmatrix} \mu_{S_{1i}} \\ \mu_{S_{2i}} \\ \beta_i \\ \alpha_i \end{pmatrix} = \begin{pmatrix} \mu_{S_1} \\ \mu_{S_2} \\ \beta \\ \alpha \end{pmatrix} + \begin{pmatrix} m_{S_{1i}} \\ m_{S_{2i}} \\ b_i \\ a_i \end{pmatrix} \tag{11}$$

where the second term on the right hand side of (11) is assumed to follow a zero-mean normal distribution with dispersion matrix

$$D = \begin{pmatrix} d_{S_1S_1} & d_{S_1S_2} & d_{S_1b} & d_{S_1a} \\ d_{S_2S_1} & d_{S_2S_2} & d_{S_2b} & d_{S_2a} \\ d_{bS_1} & d_{bS_2} & d_{bb} & d_{ba} \\ d_{aS_1} & d_{aS_2} & d_{ab} & d_{aa} \end{pmatrix}.$$

Hence a linear mixed model results. When the efects in (11) are assumed to be fixed, then a so-called fixed-effects model follows. The setting described above naturally lends itself for the validation of two scales at both the trial level as well as the individual level.

3.4.1 Trial-Level Surrogacy

In order to investigate the trial-level concurrent and/or predictive validity of two psychiatric scales, it is of interest to investigate how a change in treatment effect on one measurement scale can be translated into the other psychiatric measurement instrument. Therefore, it is essential to explore the quality of the prediction of the treatment effect on S_1 in trial i by (a) information obtained in the validation process based on trials $i = 1, \ldots, N$, and (b) the estimate of the effect of Z on S_2 in a new trial $i = 0$. Whenever there is no clear standard but simply relations are studied, as is often the case with psychometric instruments, the reverse prediction (on S_2 based on the effect on S_1) is also important.

To this end, observe that $(\beta + b_0 | m_{S10}, a_0)$ follows a normal distribution with mean and variance

$$E(\beta + b_0 | m_{S_20}, a_0) = \beta + \begin{pmatrix} d_{S_2b} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{S_2S_2} & d_{S_2a} \\ d_{S_2a} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} \mu_{S_20} - \mu_{S_2} \\ \alpha_0 - \alpha \end{pmatrix}, \tag{12}$$

$$\mathrm{Var}(\beta + b_0 | m_{S_20}, a_0) = d_{bb} - \begin{pmatrix} d_{S_2b} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{S_2S_2} & d_{S_2a} \\ d_{S_2a} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{S_2b} \\ d_{ab} \end{pmatrix}. \tag{13}$$

Similarly, $(\alpha + a_0|m_{S0}, \alpha_0)$ follows a normal distribution with mean and variance:

$$E(\alpha + a_0|m_{S_10}, b_0) = \alpha + \begin{pmatrix} d_{S_1a} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{S_1S_1} & d_{S_1b} \\ d_{S_1b} & d_{bb} \end{pmatrix}^{-1} \begin{pmatrix} \mu_{S_10} - \mu_{S_1} \\ \beta_0 - \beta \end{pmatrix}, \quad (14)$$

$$\mathrm{Var}(\alpha + a_0|m_{S_10}, b_0) = d_{aa} - \begin{pmatrix} d_{S_1a} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{S_1S_1} & d_{S_1b} \\ d_{S_1b} & d_{bb} \end{pmatrix}^{-1} \begin{pmatrix} d_{S_1a} \\ d_{ab} \end{pmatrix}. \quad (15)$$

To assess the validity of S_2 with respect to S_1 we propose to follow the suggestion of Buyse(17) *et al.* and look at the coefficient of determination:

$$R^2_{trial(f)} = R^2_{b_i|m_{S_2i},a_i} = \frac{1}{d_{bb}} \begin{pmatrix} d_{S_2b} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{S_2S_2} & d_{S_2a} \\ d_{S_2a} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{S_2b} \\ d_{ab} \end{pmatrix}. \quad (16)$$

Again, when none of the two scales can be assumed to be a standard, we may also have to look at the second coefficient of determination:

$$R^2_{trial(f)} = R^2_{a_i|m_{S_1i},b_i} = \frac{1}{d_{aa}} \begin{pmatrix} d_{S_1a} \\ d_{ab} \end{pmatrix}^T \begin{pmatrix} d_{S_1S_1} & d_{S_1b} \\ d_{S_1b} & d_{bb} \end{pmatrix}^{-1} \begin{pmatrix} d_{S_1a} \\ d_{ab} \end{pmatrix}. \quad (17)$$

These coefficients are unitless and range in the unit interval, two desirable features for interpretation. Whenever these quantities are sufficiently close to 1, we can say that one scale is a good surrogate for the other at trial level.

An attractive special case of (16) applies when the prediction of the treatment effect can be done independently of the trial-specific random intercept m_{S0}. In that case formulas (12)-(15) respectively reduce to:

$$E(\beta + b_0|a_0) = \beta + \frac{d_{ab}}{d_{aa}}(\alpha_0 - \alpha), \quad (18)$$

$$\mathrm{Var}(\beta + b_0|a_0) = d_{bb} - \frac{d^2_{ab}}{d_{aa}}, \quad (19)$$

$$E(\alpha + a_0|b_0) = \alpha + \frac{d_{ab}}{d_{bb}}(\beta_0 - \beta), \quad (20)$$

$$\mathrm{Var}(\alpha + a_0|b_0) = d_{aa} - \frac{d^2_{ab}}{d_{bb}}, \quad (21)$$

leading to a simplified coefficient of determination

$$R^2_{trial(r)} = R^2_{b_i|a_i} = \frac{d^2_{ab}}{d_{aa}d_{bb}}, \quad (22)$$

which is now symmetric on both scales. Clearly this is a very attractive property when validating two psychiometric scales for which in many cases no gold standard can be assigned. In contrast to previous approaches only one quantity suffices to assess the validity.

3.4.2 Individual-Level Surrogacy

To validate two scales at the individual level, we follow the suggestion by Buyse(17) *et al.* and consider the squared correlation between the two instruments after adjustment for both the trial effects as well as the treatment effect:

$$R^2_{indiv} = R^2_{\varepsilon_{S_{1i}}|\varepsilon_{S_{2i}}} = \frac{\sigma^2_{S_1S_2}}{\sigma_{S_1S_1}\sigma_{S_2S_2}}. \tag{23}$$

4 DATA ANALYSES

4.1 A Meta-analysis of Trials in Schizophrenic Subjects

In this section we will apply the methods of Section 3 to the data described in Section 2.1. Evidently, there is no natural "true endpoint" associated with these kind of data. Nevertheless, we will show how these methods can be used to investigate the criterion validity between the three scales of interest: PANSS, BPRS and CGI. We will successively consider the relationships between (i) PANSS and BPRS (Section 4.1.1), (ii) PANSS and CGI (Section 4.1.2) and (iii) BPRS and CGI (Section 4.1.3). Within each of these subsections, missing values (if any) were deleted first. The binary indicator for treatment (Z_{ij}) will be set to 0 for the conventional antipsychotic agents and to 1 for risperidone.

4.1.1 Relationship between PANSS and BPRS

The relationship between PANSS and BPRS was studied first. Since the BPRS is essentially constructed from the PANSS by selecting some of its items, there is a natural link between these two scales but it remains difficult to assign one of the two endpoints as the "true endpoint". With our notation we assume PANSS plays the role of S_1 and BPRS plays the role of S_2. Figure 1 (a) shows a scatterplot of BPRS versus PANSS. Clearly, both scales are highly correlated. The Pearson's correlation coefficient equals $\rho = 0.96$.

Let us now apply the different validation methods, described in Section 3. Starting with the Prentice criteria, all them are fulfilled: the treatment is prognostic for both PANSS and BPRS, BPRS is prognostic for PANSS and vice-versa, and there is no effect of treatment on either scale after correction for the other scale. A summary of these results is shown in Table 1. However one has to keep the conceptual difficulties with this formalism in mind. In addition, the lack of symmetry of this approach is a further drawback. Next, we calculated Freedman's proportion explained as $PE(S_1, S_2) = 0.875$ with

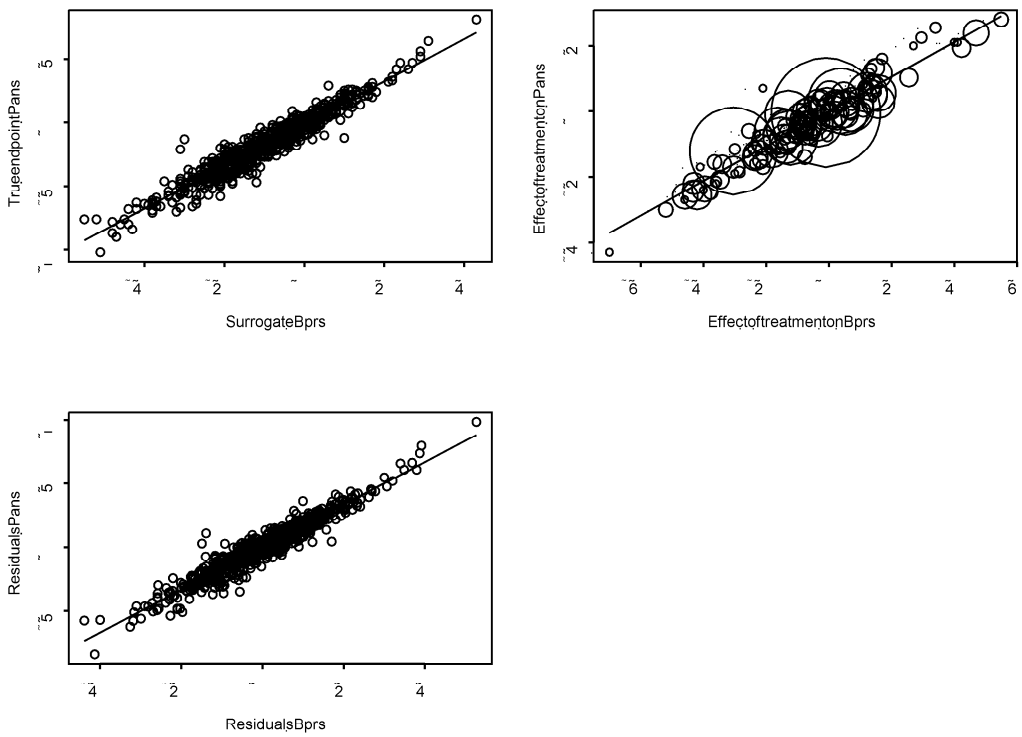

Figure 1: *(a) Scatter Plot of BPRS versus PANSS;(b) Treatment Effects on PANSS by Treatment Effects on BPRS. The size of each point is proportional to the number of patients examined by the corresponding investigator; (c) Plot of the residuals of BPRS versus PANSS.*

95% confidence interval $[0.65, 1.05]$. Because of the symmetry in the endpoints we also needed to calculate $PE(S_2, S_1) = 1.052$ with 95% confidence interval $[0.87, 1.41]$. Note that with this approach we might not only find a value of PE which is larger than 1, but in addition the confidence intervals tend to be rather wide. The relative effects and adjusted association were respectively calculated as $RE(S_1, S_2) = 1.90$ with 95% confidence interval $[0.70, 5.77]$, $RE(S_2, S_1) = 1/RE(S_1, S_2) = 0.53$ with 95% confidence interval $[0.17, 1.43]$ and $\gamma_Z = 0.96$ with confidence interval $[0.95, 0.97]$. The confidence intervals around the REs may be too large to convey any useful information. In contrast, the adjusted association is very close to one and estimated with high precision. This implies that, after accounting for treatment, a very large part of the variability of BPRS can be explained by PANSS (and vice-versa) at the individual level. In addition, one can observe the closeness with the Pearson's correlation coefficient ρ which is traditionally calculated to investigate the concurrent validity between two psychometric rating scales.

Prentice Criteria	Parameter estimated (standard error)		p-value
(2)	−4.63	(1.65)	0.0051
(3)	−2.43	(0.95)	0.0106
(4)	1.66	(0.01)	0.000
(5)	−0.57	(0.46)	0.217
(6)	0.55	(0.01)	0.000
(7)	0.13	(0.27)	0.641

Table 1: *Prentice Criteria for the comparison of PANSS versus BPRS*

Let us now consider the multi-trial approach of Buyse(17) *et al.*, which is known to be a useful validation technique (Molenberghs(14) *et al.*). Throughout, the sample sizes of the units were used to weight the observations in the calculation of the R^2 values. Figure 1 (b) shows a plot of the treatment effects on the PANSS versus the treatment effects on the BPRS for the different units. These seem to be highly correlated. Indeed, using the multi-trial method we found high conclusive values for the coefficients of determination at the trial *and* individual level. Since no clear "true endpoint" could be assigned we calculated both $R^2_{b_i|a_i,m_{S_2}} = 0.91$ (95% confidence interval: [0.86,0.94]) and $R^2_{a_i|b_i,m_{S_1}} = 0.91$ (95% confidence interval: [0.86,0.94]). However, calculating the estimate (22) based on the reduced model we found $R^2_{b_i|a_i} = 0.92$ with 95% confidence interval $[0.91, 0.93]$, which is very close to the previous values but has the advantage of being symetric in both scales. Its value indicates that not much would be gained in the precision of the prediction if instead of the full model the reduced model were used to predict the treatment effect. The individual coefficient of determination was calculated as $R^2_{\text{indiv}} = 0.92$ with 95% confidence interval $[0.91, 0.93]$. Note that this quantity is symetric in both scales. Graphically this correlation is represented by the residual plot shown in Figure 1 (c).

4.1.2 Relationship between PANSS and CGI

As pointed out before there is no natural true endpoint associated with these kind of data. Therefore, we will study the symetric relationship between PANSS (S_2) and CGI (S_1), i.e. we will let each of the endpoints play the role of "true" endpoint. This way we will be able to study the impact of changing the role of surrogate and true endpoints on the scales.

Let us start again from the Prentice Criteria. As can be read from Table 4.1.2, all the criteria were fulfilled: the treatment is prognostic for both PANSS and CGI, PANSS is prognostic for CGI (and vice-

versa) and there is no effect of treatment on either scale after correcting for the other scale. IIowever, as

Prentice Criteria	Parameter estimated (standard error)		p-value
(2)	−0.24	(0.103)	0.016
(3)	−4.46	(1.656)	0.007
(4)	0.04	(0.001)	0.000
(5)	−0.04	(0.071)	0.513
(6)	11.66	(0.402)	0.000
(7)	−1.59	(1.152)	0.167

Table 2: *Prentice Criteria for the comparison of PANSS versus CGI*

pointed out by Buyse(12) and Molenberghs and Buyse(17) *et al.*, one has to be very careful in interpreting these results, since Prentice's Criteria are surrounded with a number of conceptual difficulties, possibly leading to wrong conclusions. The point estimates for Freedman's proportions explained were estimated as $PE(S_1, S_2) = 0.81$ (95% confidence interval $[0.46, 1.67]$) and $PE(S_2, S_1) = 0.64$ (95% confidence inteval $[0.31, 1.12]$). Clearly the confidence intervals are too wide to be informative. In addition, the upper bounds again exceed 1, which is hard to justify for a proportion. The estimated values for the relative effect were $RE(S_1, S_2) = 0.055$ with 95% confidence interval $[0.01, 0.16]$ and $RE(S_2, S_1) = 18.07$ with 95% confidence interval $[6.24, 61.93]$. The treatment-adjusted association had an estimated value of $\gamma_Z = 0.72$ with confidence interval $[0.69, 0.75]$. While the point estimate for γ_Z is smaller than in the previous case, which is not so surprising given the nature of the data, it is still estimated with high precision (in contrast to the RE measures). The meta-analytic approach yielded $R^2_{b_i|m_{S_i},a_i} = 0.56$ (95% confidence interval $[0.43, 0.68]$), $R^2_{a_i|m_{T_i},b_i} = 0.56$ (95% confidence interval $[0.43, 0.68]$) at the trial level and $R^2_{\text{indiv}} = 0.51$ with 95% confidence interval $[0.47, 0.55]$ at the individual level. Clearly, these quantities were estimated with sufficient precision, at the same time indicating that the agreement between PANSS and CGI, is smaller than would have been anticipated from the classical validation approaches such as the Prentice criteria and the proportion explained. The individual level correlation between the two endpoints is relatively strong with a value of 0.71 and a 95% confidence interval, $[0.68, 0.74]$. This agrees closely with the treatment-adjusted association parameter γ_Z and even the Pearson's correlation coefficient $\rho = 0.73$. Figure 2(a) and (b) respectively show a scatterplot of CGI versus PANSS and a plot of the treatment effects on CGI by the treatment effects on PANSS, the latter being a graphical representation of R_{trial}. The R_{indiv} is graphically represented by the residual plot in Figure 2(c) Clearly, these effects are less correlated than in the previous section. In addition we calculated the R^2 measure at the trial level for the "reduced"

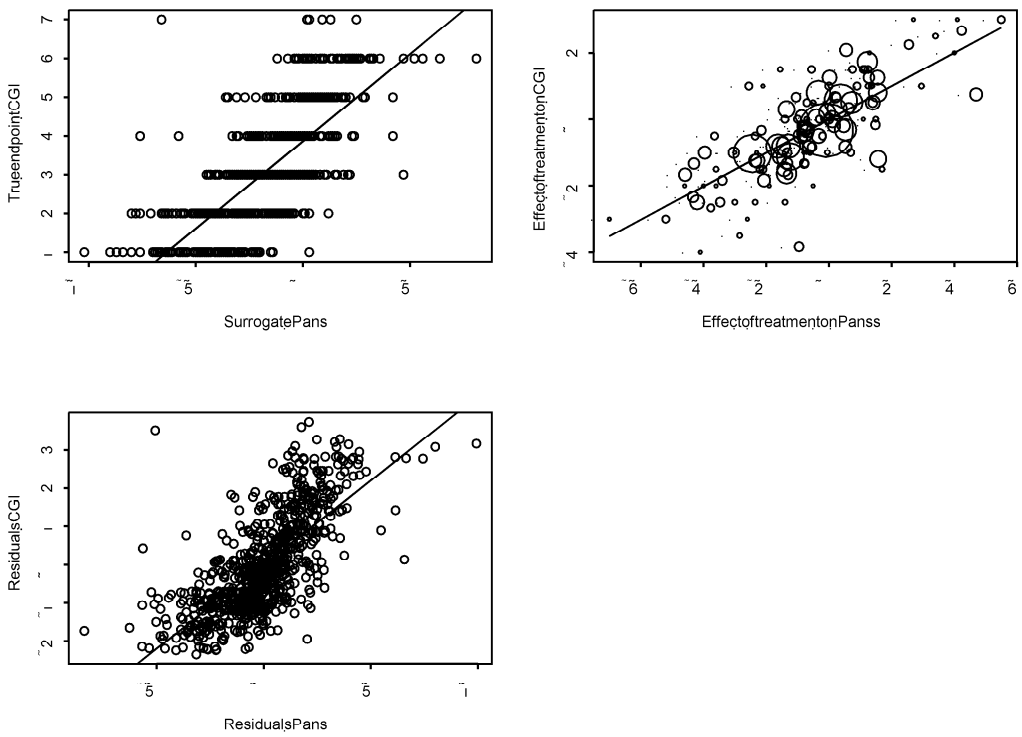

Figure 2: *Treatment Effects on CGI by Treatment Effects on PANSS. The size of each point is proportional to the number of patients examined by the corresponding investigator.*

model. This yielded $R^2_{b_i|a_i} = R^2_{a_i|b_i} = 0.56$ with 95% confidence interval $[0.43, 0.67]$ which coincides with the trial-level values obtained from the "full" model. Apart from the attractive feature that this quantity is symetric in both scales, the result again indicates that not much would be gained in the precision of the treatment prediction if instead of the full model, the reduced model were used.

In the above meta-analytic analyses we used the investigator as unit of analysis. As pointed out in Section 2.1, this lead to a total of 138 units with the number of patients per unit ranging from 2 to 30. Table 3 shows the frequency table of the number of units with a given number of patients. Clearly, the majority of units consists of less than 5 patients. Alternatively, one could also consider the main investigator as unit of analysis. For 4 out of the 5 trials only one main investigator was used leading to extremely large investigator sites. This lead to a total number of 29 units with the number of patients per unit ranging from 4 to 450, 4 of which represent trials. When redoing the meta-analytic approach for this setting we found similar results as before (we now only look at the reduced model only); the trial level and

nr. (n) of patients per unit	nr. of units with n patients	nr. (n) of patients per unit	nr. of units with n patients
2	29	10	2
3	18	11	4
4	23	12	2
5	16	13	3
6	9	15	1
7	12	18	1
8	10	21	1
9	6	30	1

Table 3: *Frequency Table of the Number of Units with a Given Number of Patients*

individual level assocation measures are respectively given by $R^2_{\text{trial(r)}} = 0.58$ (95% confidence interval [0.45,0.71]) and $R^2_{\text{indiv(r)}} = 0.52$ (95% confidence interval [0.48,0.56]). While the point estimates of these R^2 values are similar to the ones found in the previous setting, the confidence interval for R^2_{trial} is much wider, probably due to the lesser amount of trials.

Based on the results of the above meta-analytic method, we are able to predict for example the treatment effect on the CGI response based on the observed treatment effect on PANSS (or vice versa). The details hereof have been described in Section 3, equations (18) and (19). Table 4 reports prediction intervals for the 29 units together with the number of patients per unit. In this table, $\hat{\alpha_0}$ and $\widehat{\beta + b_0}$ are values estimated from the data; $E(\beta + b_0)$ is the predicted treatment effect on CGI, given its effect on PANSS. Clearly, in all cases, the predicted values for $\beta + b_0$ agree reasonably well with the effects estimated from the data.

An interesting plot is shown in Figure 3 which indicates how effect changes on one outcome can be translated on another outcome. Translating effect changes of PANSS or BPRS to the CGI scale is more or less similar. But, as expected, the translation of an effect change on BPRS to PANSS is much more precise.

4.1.3 Relationship between BPRS and CGI

When studying the relationship between CGI (S_1) and BPRS (S_2) we found similar results to the ones obtained in Section 4.1.2. This is not so surprising given the strong relationship found between BPRS and PANSS. Since results for the full and reduced models almost coincide, we only present the values for the reduced model here.

Unit	# patients	$\hat{\alpha_0}$	$E(\beta + b_0 \mid a_0)$	$\widehat{\beta + b_0}$
1	8	14.00 (16.35)	0.53 (0.63)	0.50 (1.26)
2	6	-43.33 (29.02)	-1.99 (0.63)	-2.33 (1.25)
3	9	-13.50 (12.75)	-0.75 (0.60)	0.30 (1.18)
4	4	7.50 (35.28)	0.08 (0.58)	1.50 (1.80)
5	9	-7.60 (7.65)	-0.45 (0.63)	-0.40 (0.99)
6	8	-42.00 (18.93)	-1.88 (0.63)	-2.50 (1.04)
7	7	-39.58 (18.71)	-2.07 (0.61)	-1.00 (1.18)
8	6	-13.33 (13.79)	-0.69 (0.62)	-1.33 (1.56)
9	6	-7.33 (23.35)	-0.44 (0.63)	-0.33 (1.33)
10	4	-2.00 (18.06)	-0.18 (0.63)	-0.50 (1.80)
11	68	-4.84 (4.46)	-0.32 (0.63)	-0.47 (0.36)
12	8	-14.25 (30.53)	-0.72 (0.62)	-1.50 (0.89
13	7	-6.33 (11.24)	-0.37 (0.63)	-0.83 (0.95)
14	4	-36.5 (14.77)	-1.96 (0.58)	-0.50 (0.50)
15	5	-13.00 (26.93)	-0.66 (0.61)	-1.66 (1.72)
16	8	-22.75 (10.45)	-1.13 (0.63)	-1.25 (0.63)
17	8	-9.00 (10.93)	-0.52 (0.63)	-0.50 (0.65)
18	450	-3.57 (2.13)	-0.28 (0.63)	-0.15 (0.13)
19	7	-23.5 (12.02)	-1.16 (0.63)	-1.25 (0.74)
20	5	-5.33 (13.52)	-0.33 (0.63)	-0.83 (0.57)
21	70	2.75 (5.79)	-0.00 (0.63)	0.21 (0.38)
22	7	-7.50 (16.13)	-0.46 (0.63)	-0.25 (1.40)
23	7	-20.66 (15.39)	-1.00 (0.62)	-1.83 (1.06)
24	9	-4.00 (11.06)	-0.31 (0.63)	0.05 (0.93)
25	5	-7.83 (11.16)	-0.43 (0.61)	-1.33 (0.86)
26	45	-20.15 (9.68)	-1.01 (0.63)	-1.18 (0.50)
27	9	1.14 (19.19)	-0.06 (0.63)	0.00 (0.95)
28	5	-10.50 (10.96)	-0.63 (0.59)	0.66 (0.86)
29	8	-3.25 (10.71)	-0.24 (0.63)	-0.49 (0.79)

Table 4: *Predictions for the treatment effects on CGI based on the observed treatment effects on PANSS. Estimates (standard errors) are shown.*

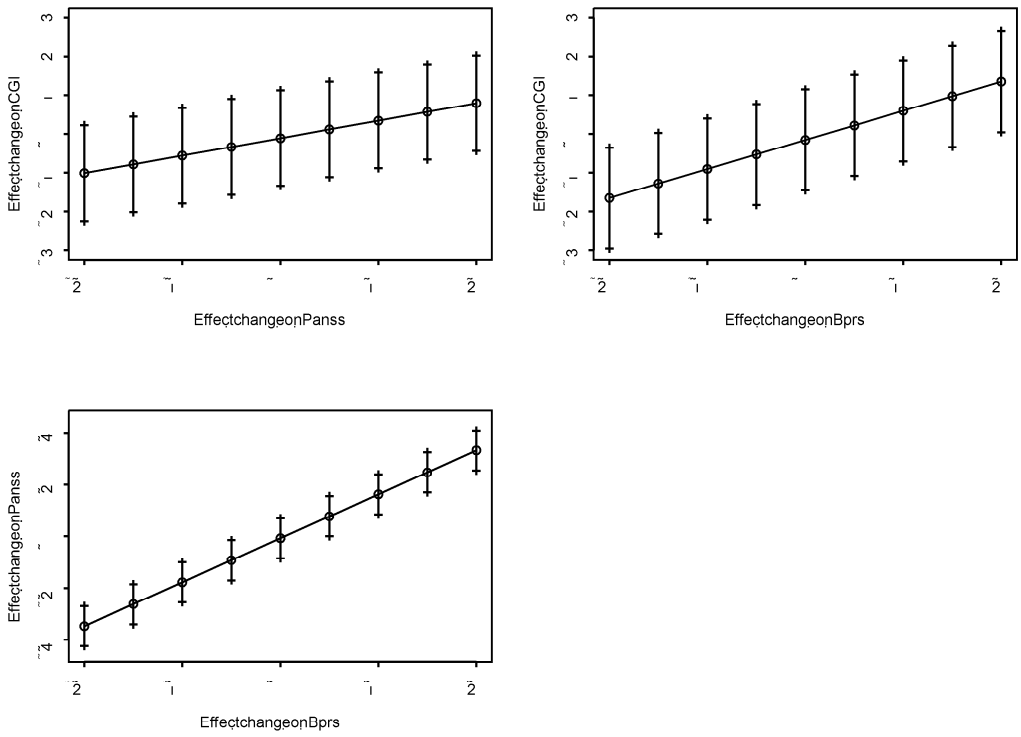

Figure 3: *Effect Changes on one Outcome by the Effect Changes on Another Outcome.*

Again, the Prentice criteria were fulfilled as can be seen from the summary presented in Table 4.1.3: Freedman's proportions explained were estimated as $PE(S_1, S_2) = 0.72$ with a wide 95% confidence interval of $[0.37, 1.49]$ and $PE(S_2, S_1) = 0.09$ with 95% confidence interval of $[0.33, 1.34]$. The estimated value for the relative effect $RE(S_1, S_2)$ is 0.10 with 95% confidence interval $[0.03, 0.34]$ and the treatment-adjusted association has an estimated value of $\gamma_Z = 0.71$ with confidence interval $[0.68, 0.73]$. Using the meta-analytic approach we find a value of 0.59 for R^2_{trial} with 95% confidence interval $[0.46, 0.73]$ and $R^2_{\text{indiv}} = 0.49$ with 95% confidence interval $[0.44, 0.53]$. Figure 4 (a)-(c), as before, show respectively the scatterplot of CGI versus BPRS, the treatment effects on CGI by the treatment effects on BPRS and a residual plot.

4.2 An Equivalence Trial in Schizophrenic Patients

In Section 3 we presented a brief history of different validation techniques that have been so far proposed for surrogate markers. A thorough study of the available literature shows how the classical techniques such

Prentice Criteria	Parameter estimated (standard error)	p-value
(2)	-0.24 (0.103)	0.016
(3)	-2.35 (0.954)	0.013
(4)	0.07 (0.002)	0.000
(5)	-0.06 (0.072)	0.363
(6)	6.62(0.235)	0.000
(7)	-0.73(0.673)	0.279

Table 5: *Prentice Criteria for the comparison of BPRS versus CGI*

as Prentice's criteria and the proportion explained, but also the relative effect and adjusted association are surrounded by difficulties. Of course, this evolution has been indispensable in the development of new insights and formalizing new validation approaches.

The present section illustrates on the basis of the data described in Section 2.2 how the classical approaches can hide the possible "agreement" of variables in an equivalence study and how they can produce misleading or even wrong results. Like in the previous section we will subsequently consider the relationships between (i) PANSS and BPRS (Section 4.3) and (ii) PANSS and CGI (Section 4.3.1). Results about the BPRS versus CGI agreement are not shown. They are very similar to the results obtained in Section 4.3.1.

4.3 PANSS versus BPRS

Just for sake of illustration we let PANSS play the role of "true" endpoint. The Prentice criteria now utterly failed to show the high agreement between both scales. Results are summarized in Table 6. By definition of an equivalence trial, the first two criteria are bound to be unfulfilled.

As always, Freedman's proportion explained cannot give a conclusive answer, being estimated at $PE = -0.525$ with an infinite 95% confidence interval. Apart from the confidence interval which is too wide to be of any practical use, the PE is even negative which can hardly be justified for a proportion and makes it hard to interpret. The relative effect was estimated at $RE = -3.14$ with an unbounded confidence interval as well, which makes it inconclusive. However the adjusted association equals $\gamma_Z = 0.97$ with confidence interval $[0.97, 0.98]$, giving evidence of a high individual level assocation corrected for treatment. The meta-analytic approach produced values, $R^2_{\text{trial(r)}} = 0.96$ with 95% confidence interval $[0.82, 1.09]$ at the

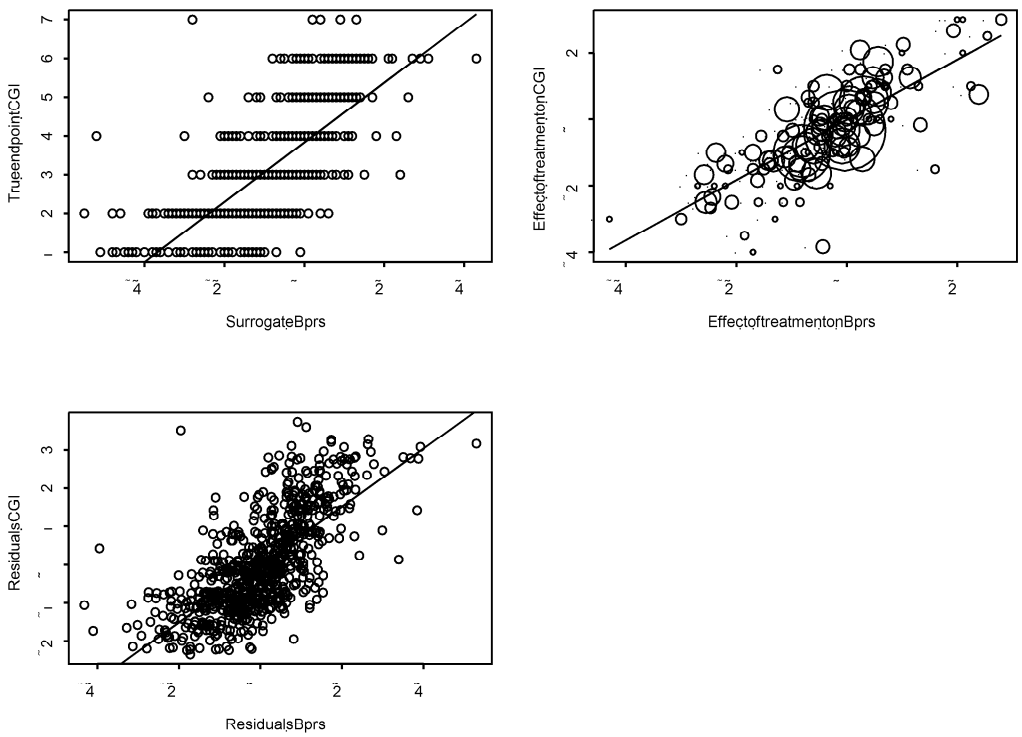

Figure 4: *Treatment Effects on CGI by Treatment Effects on BPRS. The size of each point is proportional to the number of patients examined by the corresponding investigator.*

trial level, and $R^2_{\text{indiv(r)}} = 0.94$ with 95% confidence interval $[0.92, 0.95]$ at the individual level. Both give conclusive results, which are in agreement with the ones found in Section 4.1.1. This "robust" behaviour clearly confirms the superiority of the meta-analytic approach. Thus, we have illustrated the meta-analytic approach is the only that is able to use data from equivalence trials for validation. All other approaches give inconclusive results, with the Prentice criteria being even utterly useless by definition.

4.3.1 PANSS versus CGI

Let us now investigate the agreement between PANSS and CGI with CGI playing the role of "true" or "standard" endpoint. A summary of the Prentice criteria is found in Table 7. As could have been anticipated, the first two criteria are again not fulfilled. Freedman's proportion explained takes a negative value of $PE = -0.94$ with an infinite confidence interval. The relative effect estimate was estimated at $RE = -0.03$ with also an infinite confidence interval. The adjusted association was estimated as $\gamma_Z = 0.74$

Prentice Criteria	Parameter estimated (standard error)		p-value
(2)	1.06	(4.050)	0.792
(3)	−0.33	(2.398)	0.887
(4)	1.65	(0.024)	0.000
(5)	1.62	(0.834)	0.052

Table 6: *Prentice Criteria for the comparison of PANSS versus BPRS*

with confidence interval $[0.69, 0.79]$, which closely corresponds to the value obtained in Section 4.1.2. The meta-analytic approach yielded values, $R^2_{\text{trial(r)}} = 0.70$ with 95% confidence interval $[0.44, 0.96]$ at the trial level, and $R^2_{\text{indiv(r)}} = 0.55$ with 95% confidence interval $[0.47, 0.62]$ at the individual level. This illustrates again that the multi-trial approach is the only one that seems to give conclusive results, which are consistent with the ones found in Section 4.1.2.

5 DISCUSSION

In this paper we have shown how a well-known psychometric property such as the criterion validity can be assessed using techniques that have been recently developed in the field of surrogate marker validation in clinical trials. While psychiatric studies, such as the ones presented here, differ from clinicical trials by the fact that no true endpoint can be assigned, we show that the developed methodology can equally well be applied on softer endpoints.

Traditional psychometric techniques that try to assess the criterion validity are often limited to the calculation of simple Pearson correlation coefficients. In contrast, the multi-trial approach described in this paper allows us to relate or predict a treatment effect on one scale with a treatment effect on the other scale. Further, one is able to distinguish between trial-level and individual-level agreement, which the classical techniques do not. In addition, treatment effects on aggregate scores can be translated to effects on more understandable measures.

While we have looked at validation techniques in cross-sectional studies only, it would be of interest to construct multi-trial techniques when both outcomes have repeated measurements of time. This is the subject of ongoing research.

Prentice Criteria	Parameter estimated (standard error)		p-value
(2)	−0.03	(1.186)	0.835
(3)	1.06	(4.050)	0.792
(4)	0.03	(0.002)	0.000
(5)	−0.07	(0.124)	0.544

Table 7: *Prentice Criteria for the comparison of PANSS versus CGI*

Acknowledgements

The first author gratefully acknowledges support from an LUC Bijzonder Onderzoeksfonds grant. The second author was supported by the Institute for the Promotion of Innovation by Science and Technology (IWT) in Flanders, Belgium. The authors are also grateful to Janssens Pharmaceutica for permission to use their data. Research supported by a PAI program P5/24 of the Belgian Federal Government (Federal Office for Scientific, Technical, and Cultural Affairs).

REFERENCES

1. Streiner DL, Norman GR: *Health Measurement Scales: a Practical Guide to their Developemnt and Use*, Oxford University Press, 1995.

2. Cronbach LJ: "Coefficient Alpha and the Internal Structure Tests," *Psychometrika*, **51**, 297–334, 1951.

3. Kuder GF, Richardson MW: "The Theory of Estimation of Test Reliability", *Psychometrika*, **2**, 151–160, 1953.

4. Fleiss J, Cohen J: "The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability," *Educational and psychological measurement*, **33**, 6113–6119, 1973.

5. Deyo RA, Dierh P, Patrick D: "Reproducibility and Responsiveness of Health Status Measure Statistics and Strategies for Evaluation," *Controlled Clinical Trials*, **12**, 142–158, 1991.

6. Campbell DT, Fisk DW: "Convergent and Discriminant Validation by the Multitrait Multi-method Matrix", *Psychological Bulletin*, **56**, 85–105, 1959.

7. Boissel JP, Collet JP, Moleur P, IIaugh M: "Surrogate endpoints: a basis for a rational approach," *European Journal of Clinical Pharmacology*, **43**, 235–244, 1992.

8. Fleming TR, DeMets DL: "Surrogate endpoints in clinical trials: are we being misled ?," *Annals of Internal Medicine* **125**, 605–613, 1996.

9. De Gruttola V, Fleming TR, Lin DY, Coombs R: "Validating surrogate markers - are we being naive ?," *Journal of Infectious Diseases* **175**, 237–246, 1997.

10. Prentice RL: "Surrogate endpoints in clinical trials: definitions and operational criteria," *Statistics in Medicine* **8**, 431–440, 1989.

11. Fleming TR, Prentice RL, Pepe MS, Glidden, D: "Surrogate and Auxiliary Endpoints in Clinical Trials, with Potential Applications in Cancer and AIDS research," *Statistics in Medicine*, **13**, 955–968, 1994.

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

13. Freedman LS, Graubard BI, Schatzkin A: "Statistical validation of intermediate endpoints for chronic diseases," *Statistics in Medicine* **11**, 167–178, 1992.

14. Molenberghs G, Buyse M, Burzykowski T, Renard D, Geys H: "Statistical challenges in the evaluation of surrogate endpoints in randomized trials," *Submitted for publication*, 2000.

15. Albert JM, Ioannidis JPA, Reichelderfer P, Conway B, Coombs RW, Crane L, Demasi R, Dixon DO, Flandre P, Hughes MD, Kalish LA, Larntz K, Lin D, Marschner IC, Munoz A, Murray J, Neaton J, Pettinelli C, Rida W, Taylor JMG, Welles SL: "Statistical Issues for HIV surrogate endpoints: point/counterpoint," *Statistics in Medicine,* **17**, 2435–2462, 1998.

16. Daniels MJ, Hughes MD: "Meta-Analysis for the Evaluation of Potential Surrogate Markers," *Statistics in Medicine*, 16, 1515–1527, 1997.

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

18. Gail MH, Pfeiffer R, Van Houwelingen HC, Carroll RJ: "On meta-analytic assessment of surrogate endpoints," *Biostatistics*, **1**, 231–246, 2000.

19. Verbeke G, Molenberghs G: *Lecture Notes in Statistics. Linear Mixed Models in Practice: A SAS-Oriented Approach*, New York: Springer, 1997.

20. Geys H: *Pseudo-likelihood Methods and Generalized Estimating Equations: Efficient Estimation Techniques for the Analysis of Correlated Multivariate Data*, unpublished Phd thesis, 1999.

21. Molenberghs G, Geys H, Buyse M: "Mixed Discrete and Continuous Outcomes for the Validation of Surrogate Endpoints in Randomized Experiments," *Statistics in Medicine*, **00**, 000–000, 2001.

22. Renard D, Geys H, Molenberghs G, Burzykowski T, Buyse M: "Validation of surrogate endpoints in multiple randomized clinical trials with discrete endpoints," submitted, 2001.

23. Burzykowski T, Molenberghs G, Buyse M, Geys H, Renard D: "Validation of surrogate endpoints in multiple randomized clinical trials with failure-time endpoints," *Applied Statistics*, **00**, 000–000, 2001.

24. Kay SR, Fiszbein A, Opler LA: "The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia", *Schizophrenia Bulletin*, **13**, 261–276, 1987.

25. Kay SR, Opler LA, Lindenmayer JP: "Reliability and Validity of the Positive and Negative Syndrome Scale for Schizophrenics," *Psychiat. Res*, **23**, 99–110, 1988.

26. Overall JE, Gorham DR: "The Brief Psychiatric Rating Scale", *Psychol. Rep.*, **10**, 799 812, 1962.

27. Peuskens J and the Risperidone Study Group: "Risperidone in the treatment of chronic schizophrenic patients: a multinational, multicentre, double-blind, parallel-group study versus haloperidol," *Br J Psychiatry*, **166**, 712–726, 1995.

28. Chounard G, Jones B, Remington G: "A canadian multicenter placebo-controlled study of fixed doses of risperidone and haloperidol in the treatment of chronic schizophrenic patients," *J. Clin. Phychopharmacol*, **13**, 25–40, 1993.

29. Marder SR, Meibach RC: "Risperidone in the treatment of schizophrenia," *Am. J. Psychiatry*, **151**, 825–835, 1994.

30. Hoyberg OJ, Fensbo C, Remvig J, Lingjaerde OK, Slotei-Nielsen M, Salvesen I: "Risperidone versus perphenazine in the treatment of chronic schizophrenic patients with acute exacerbations," *Acta Psychiatr Scand*, **88**, 395–402, 1993.

31. Blin O, Azorin JM, Bouhours P: "Antipsychotic and anxiolytic properties of risperidone, haloperidol and methotrimeprazine in schizophrenic patients," *J. Clin. Psychopharmacol*, **16**, 38–44, 1996.

32. Huttunen MO, Piepponen T, Rantanen H, Larmo I, Nyholm R, Raitasuo V: "Risperidone versus zuclopenthixol in the treatment of acute schizophrenic episodes: a double-blind parallel-group trial," *Acta Psychiatr Scand*, **91**, 271–277, 1995.

33. Nair NPV and the Risperidone Study Group: "Therapeutic equivalence of risperidone given once daily and twice daily in patients with schizophrenia," *Journal of Clinical Psychopharmacology*, **18**, 103–110, 1998.

34. Lin DY, Fleming TR, De Gruttola V: "Estimating the Proportion of Treatment Effect Explained by a Surrogate Marker," *Statistics in Medicine,* **16**, 1515–1527, 1997.

35. Flandre P, Saidi Y: "Letters to the editor: Estimating the proportion of treatment effect explained by a surrogate marker," *Statistics in Medicine* **18**, 107–115, 1999.

36. Buyse M, Molenberghs G, Burzykowski T, Renard D and Geys H: "Statistical Validation of Surrogate Endpoints: Problems and Proposals," *Drug Information Journal*, **34**, 557-454, 2000.