Using longitudinal data from a clinical trial in depression to assess the reliability of its outcome scales

LAENEN, Annouschka; ALONSO ABAD, Ariel; MOLENBERGHS, Geert; Mallinckrodt, Craig H. & VANGENEUGDEN, Tony (2009) Using longitudinal data from a clinical trial in depression to assess the reliability of its outcome scales. In: JOURNAL OF PSYCHIATRIC RESEARCH, 43(7). p. 730-738.

DOI: 10.1016/j.jpsychires.2008.09.010
Handle: http://hdl.handle.net/1942/9203

Using longitudinal data from a clinical trial in depression to assess the reliability of its outcome scales.

Annouschka Laenen
Hasselt University, Center for Statistics,
Agoralaan 1, B3590 Diepenbeek, Belgium.
phone: +3211268292
fax: +3211268299
email: annouschka.laenen@uhasselt.be

Ariel Alonso
Hasselt University, Center for Statistics,
Agoralaan 1, B3590 Diepenbeek, Belgium.

Geert Molenberghs
Hasselt University, Center for Statistics,
Agoralaan 1, B3590 Diepenbeek, Belgium.
Katholieke Universiteit Leuven, Biostatistical Centre,
Kapucijnenvoer 35, B3000 Leuven, Belgium.

Tony Vangeneugden
Tibotec, Johnson & Johnson,
Generaal De Wittelaan L11 B3,
B2800 Mechelen, Belgium.

Craig H. Mallinckrodt
Eli Lilly & Company, Lilly Corporate Center,
Indianapolis, IN 46285, U.S.A.

Abstract

Longitudinal studies are permeating clinical trials in psychiatry. Additionally, in the same field, rating scales are frequently used to evaluate the status of the patients and the efficacy of new therapeutic procedures. Therefore, it is of utmost importance to study the psychometric properties of these instruments within a longitudinal framework. In the area of depression, the Hamilton Depression Rating Scale (HAMD) is regularly used for antidepressant treatment evaluation. However, the use of HAMD has not been exempted from criticism what has lead to the development of new scales that are expected to be more sensitive for change, such as the Montgomery-Åsberg Depression Rating Scale (MADRS). In general, the reliability of these scales has been extensively studied by using classical methods for reliability estimation, developed for specifically designed reliability studies. Unfortunately, the settings customarily considered in these reliability studies are usually far from the practical conditions in which these scales are applied in clinical trials and practice. In the present paper we assess the reliability of these instruments in a more realistic scenario thereby using longitudinal data coming from clinical studies. Nowadays, newly developed methodology based on an extended concept of reliability, allow us to use longitudinal data for reliability estimation. This new approach not only enables to avoid bias by offering a better control of disturbing factors but it also produces more precise estimates by taken advantage of the large samples sizes available in clinical trials. Further, it offers practical guidelines for an optimal use of a rating scale in order to achieve a particular level of reliability. The merits of this new approach are illustrated by applying it on two clinical trials in depression to assess the reliability of the three outcome scales, HAMD, MADRS, and the Hamilton Anxiety Rating scale (HAMA).

Keywords: Clinical trials, Depression, Longitudinal data, Rating scales, Reliability

1 Introduction

The Hamilton Rating Scale for Depression (HAMD) was developed in the late 1950s to assess the effectiveness of the first generations of antidepressants. The scale quickly became the standard measure of depression severity for the evaluation of new anti-depressive drugs and is hitherto the most commonly used measure for depression. The original rating form included 21 items, although Hamilton (1960) indicated that only 17 should contribute to the total scale score because 1 of the last 4 items represented depressive type rather than depression severity, and three others did not occur sufficiently frequently. Nine of the 17 items are rated from 0 to 4, whereas 8 items are rated 0 to 2. Concurrently, Hamilton (1959) developed one of the first rating scales to quantify the severity of anxiety symptomatology (HAMA). Several conceptual and psychometric problems with the HAMD have been described in the literature. However, reliability studies have mostly indicated satisfactory results. Bagby *et al* (2004) selected 70 studies that examined the psychometric properties of HAMD and found test-retest reliabilities ranging from 0.81 to 0.98, based on the Pearson correlation coefficient. An almost identical range of pearson correlations was found for inter-rater reliability. The Montgomery-Åsberg Depression Rating Scale (MADRS) was designed to address the limitations of the HAMD, and was supposed to measure contemporary definitions of depression and to be more sensitive to change (Montgomery and Åsberg 1979). Maier *et al* (1988) compared inter-rater reliabilities for the two depression scales based on three different studies. However no significant differences in the reliabilities of the HAMD and the MADRS were found in any of them.

It is widely known that reliability is not a fixed property of a rating scale, but is population dependent instead, with more homogeneous populations leading to lower reliabilities. This is reflected in wide ranges that are often reported when different

reliability studies on the same scale are compared. As a result, the reliability of a rating scale should be checked every time it is applied in a different population. Besides the heterogeneity of the population, other factors may have their influence on the reliability of a measurement as well, such as the skills or the training of the raters.

Several scholars, such as Fleiss (1987) and Lachin (2004), have stressed the fact that measurement error or low reliability can affect the results found in clinical trials. Ideally, the reliability of measurements should be checked every time a scale is going to be used in a clinical study. However, the organization of a supplementary investigation to assess reliability with additional data collection on top of the actual clinical trial, may be practically unfeasible.

In recent years clinical trials in psychiatry have been dominated by longitudinal study designs. Making several evaluations of the same patient at different time points offers a more complete information about the evolution of the patient and gives also the opportunity to evaluate the impact of the new therapeutic procedure on this evolution. Research has clearly shown that longitudinal analytic methods, including mixed-effects analyses, are well suited in this area (Gueorguieva and Krystal 2004, Mallinckrodt *et al* 2004, Leon *et al* 2006, Verbeke and Molenberghs 2000, Molenberghs and Kenward 2007). From this observation the question arises as to what extent the repeated measurements in such a trial could be adopted to study the test-retest reliability.

In the present work we argue that using clinical trial data for the appraisal of reliability can bring many advantages. However, some methodological challenges need to be taken into account. In the next section we will expand this idea when describing in more detail the main objectives of the present work.

2 Aims of the paper

Two longitudinal drug efficacy trials in depression applied the HAMD as the primary outcome measure, whereas the HAMA and the MADRS were secondary outcomes. The clinical results of these studies have been published by Mallinckrodt *et al* (2003). This paper however focusses on the reliability of the three outcome scales, and how this can be estimated based on the clinical trial data. We will give a brief overview of the specific aims of this work.

Obtaining unbiased reliability estimates for HAMD, MADRS and HAMA
In this paper a new methodology will be used that has important advantages compared to the classical methods. Indeed, classical methods to estimate reliability are based on strong assumptions. For instance, it is typically assumed that the status of the patient does not change during the study period. Note that this assumption would be extremely unrealistic in a clinical trial setting where the the drugs under study are expected to provoke a change in the patient's condition. Other strong assumptions are also necessary within the classical framework and we remit the reader to DeShon *et al* (1998) for a more detail account of some of them. When these assumptions are not met, biased reliability estimates may follow. Using more advanced methods, many of these assumptions can be relaxed, resulting in unbiased reliability estimates.

Comparing the reliability of HAMD and MADRS in a longitudinal setting
Since reliability depends on the population being measured as well as on study-specific aspects, it is important to base oneself on a single study when two different rating scales are to be compared. This procedure was followed by Maier *et al* (1988) to compare the HAMD and the MADRS on inter-rater reliability. However, since one of the aims of the MADRS is to be more sensitive to change, we perform the

comparison of the two scales in a longitudinal setting.

Studying reliability in different clinical populations

The methods used in the paper allow to study reliability based on the clinical outcome data, implying a large study sample. The advantage of this is twofold: first it allows us to estimate reliability with a higher level of precision. Second, it permits to study reliability in different clinical populations. For instance, in the present work we evaluate reliability in less and more severely depressed patients as well as between responders and non-responders.

Optimizing the use of the rating scales

In the paper we apply a method that implies a new way of looking at the concept of reliability. Interestingly, this new methodology shows that the reliability of the conclusions obtained through the use of a rating scale can depend on the way the scale is applied. Essentially, it depends on the number of evaluations carried out with the instrument. The practical implications of this finding are considerable. For instance, we can calculate the number of measurements necessary to achieve a minimum level of reliability.

Illustrating the advantages of more advanced analysis methods for reliability

Besides the specific results found for the HAMD, MADRS and HAMA, the paper illustrates the use of new techniques that can cope with several shortcomings in classical methods. At the cost of a more complex data analysis, these techniques could imply saving expenses and an increased precision in future clinical trials.

Remarkably, in spite of its more complicated methodological background this new approach leads to simple yet meaningful quantifications of reliability. Further, the results obtained have a clear clinical interpretation what can greatly increase their

appeal to the experts in the concrete field where they are used.

3 Data and Methods

3.1 Case Study

The case study data come from two clinical trials evaluating the efficacy of two antidepressants. The first study (Study 5 in Mallinckrodt *et al* 2003) is a randomized double-blind trial investigating the efficacy of duloxetine in the treatment of major depressive disorders. The primary endpoint was the HAMD_{17} total score, whereas the HAMA (14 items) and the MADRS (10 items) total scores were used as secondary endpoints. Measurements were taken at baseline and after $1, 2, 4, 6, 8$ and 10 weeks. The study contained a total of 354 patients of which 90 were assigned to the placebo group, 91 received Duloxetine (40 mg/d), 84 received Duloxetine (80 mg/d) and 89 received Paroxetine (20 mg/d). The design of the second study (Study 6 in Mallinckrodt *et al* 2003) is identical to the design of the first one, with 89 patients assigned to the placebo group, 86 received Duloxetine (40 mg/d), 91 received Duloxetine (80 mg/d) and 87 received Paroxetine (20 mg/d). The first three panels of Figure 1 show the individual patient profiles for the three different scales in trial 1.

3.2 Methodology

In the classical test theory (CTT; Lord and Novick 1968), the outcome of a test for subjects $i = 1, \ldots, n$ is modeled as

$$Y_i = T_i + \varepsilon_i,$$

where Y_i denotes the observed score, T_i is the true score and ε_i the corresponding measurement error. Assuming that the measurement errors are mutually uncorrelated as well as uncorrelated with the true scores, it follows that the total variability in the observed data can be decomposed as the variability emanating from the

differences between the subjects in the population plus the variability induced by measurement error, i.e., $\text{Var}(Y_i) = \text{Var}(T_i) + \text{Var}(\varepsilon_i)$. The reliability of a measuring instrument is then defined as the ratio of the true score variance to the observed score variance, i.e., the proportion of the total variance that is not attributed to measurement error

$$R = \frac{\text{Var}(T_i)}{\text{Var}(Y_i)} = 1 - \frac{\text{Var}(\varepsilon_i)}{\text{Var}(Y_i)}. \tag{1}$$

Essentially, R expresses the extent to which a difference in the observed scores reflects a real difference between the subjects or can be explained by just measurement error. This expression cannot be directly applied since the true score T_i is unobserved, however, under a strict set of assumptions (1) equals the correlation between two consecutive measurements. The main limitation of CTT is that it operates within a very narrow modeling framework which implies that its methods and definitions, though appealing, are based on very strong assumptions that are unrealistic in many practical situations. One of the most important attempts to extend reliability theory to more general settings came with the development of Generalizability theory (G-theory; Cronbach *et al* 1963, 1972), which is based on analysis-of-variance models with random effects, and was developed to model the multiple sources of variation present in a measurement system. Even though G-theory is based on a much more flexible modeling framework, when applied to longitudinal settings it still relies on strong and unrealistic assumptions (DeShon *et al* 1998, Laenen *et al* 2008). Let us illustrate this by explicitly stating two of the main assumptions that lie at the basis of CTT as well as G-theory and that may be problematic in test-retest reliability studies. The first assumption concerns the 'steady-state' of the patient, i.e., the assumption that the patient's condition does not evolve over time. The second assumption is related to the absence of correlated errors. Such a correlation may emanate, for instance, from intrinsic factors related to the phenomenon under study

or from extrinsic factors as the presence of a memory-effect of the rater. Violation of both assumptions will lead to an overestimation of the reliability coefficient (Smith and Luecht 1992, Bost 1995). Both problems are however related; by increasing the time period between two measurements one decreases the probability of a memory effect but the chance that a patient has changed increases, and vice versa. These issues are very difficult to control within the classical framework, even in a controlled reliability study.

In contrast, the family of linear mixed models (LMM) is especially suitable for modeling longitudinal measurements without requiring unrealistic assumptions (Laird and Ware 1982; Verbeke and Molenberghs 2000). This makes them the perfect tool to extend the concept of reliability to a longitudinal scenario (Vangeneugden *et al* 2004; Laenen *et al* 2007, 2008). The LMM can be written as

$$\boldsymbol{Y_i} = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b_i} + \boldsymbol{\varepsilon_i}, \tag{2}$$

where $\boldsymbol{Y}_i$ is the vector of repeated observed scores for subject i, for instance, the different HAMD scores for a subject taken at the six visits. The element $X_i\boldsymbol{\beta}$ summarizes the systematic changes. Using this part of the model one can account for systematic changes over time, the effect of a treatment or therapeutic procedure, the effect of other important factors as hospital, socioeconomic status, among others. The second part, $Z_i\boldsymbol{b_i}$, is called the random effect structure and allows to account for subject-specific elements that can explain the subject-specific characteristics at baseline as well as the individualized evolution of the patients over time. For example, one can model the fact that patients differ on their general initial HAMD score even after taking into account the effect of other factors, or one can model the fact that different patients have different evolutions over time even when they share common characteristics like treatment, gender or socioeconomic status. Finally, the

element $\boldsymbol{\varepsilon_i}$ indicates a vector of measurement errors, one for each measurement taken for patient i.

The above discussion shows that, unlike in CTT and G-theory, the evolution of the outcome measurements (e.g. HAMD scores) of a patient over time can now be incorporated into the model, making the steady-state assumption unnecessary (Vangeneugden *et al* 2004).

In what follows we assume a balanced study design, meaning that the measurements are taken at the same time points for all the patients. This does however not mean that we assume absence of missing data. Actually, one of the advantages of the new method is its more founded way of handling missing observations. Both, CTT as well as G-theory, assume that the missing generating mechanism is Missing Completely at Random (MCAR). On the other hand the new methodology assumes a much weaker condition: Missing at Random (MAR). The field of missing data is highly technical and complex. It is not the purpose of the present work to give a deep account of all the issues involved in the analysis of missing data. Let us just state that unlike the MCAR, the MAR assumption can be considered reasonable on many clinical trials and, therefore, our conclusions will usually be valid when missing data are present (Rubin 1976, Molenberghs and Kenward 2007).

As is obvious from the classical approach, quantifying reliability implies to account for the different sources of variability in the data. Unfortunately, one can not account for the variability of a set of longitudinal measurements $\boldsymbol{Y_i}$ using a single number but using a $p \times p$ variance-covariance matrix denoted by V, where p denotes the number of repeated measurements per patient. On the matrix diagonal one finds the variances of the measurements at the 1 to p time points, the off-diagonal elements describe the association between measurements at different time points. However,

in spite of this additional complexity, the variance-covariance matrix V of $\boldsymbol{Y_i}$ can be written similarly to the CTT setting, as:

$$V = \Sigma_D + \Sigma \tag{3}$$

with Σ_D the variance-covariance matrix related to the random effects or subject-specific characteristics and Σ the variance-covariance matrix of the measurement errors.

The LMM's ability to distinguish between different sources of variability makes it especially suitable for estimating reliability. Indeed, from (1) it follows that a distinction needs to be made between variability coming from the true scores of the subjects on the one hand, and the error variability, on the other hand. Based on this idea Laenen *et al* (2007, 2008) introduced two reliability measures, the so-called R_T and R_Λ coefficients. The two measures mainly differ with respect to the way they use to summarizing the variability captured by the variance-covariance matrices into a single number, which is the trace (denoted by the symbol tr) for R_T and the determinant (denoted by the symbol ||) for R_Λ. The coefficients are defined as

$$R_T = 1 - \frac{\text{tr}(\Sigma)}{\text{tr}(V)} \tag{4}$$

$$R_\Lambda = 1 - |\Sigma V^{-1}|. \tag{5}$$

Both measures take values between zero and one. They both reach the lower bound zero when the scale contains nothing but measurement error and the upper bound one when the instrument fully reflects the variability between the subjects. Note further that both proposals quantify the proportion of the total variability (V) not owing to measurement error (Σ), exactly as in the classical definition of reliability (1). Actually, when applied in a setting where the assumptions of CTT are met,

both measures reduce to the classical expression of reliability. Furthermore, when applied in a setting where G-theory assumptions are met, the two measures reduce to the index of dependability Φ and can be converted into the coefficient $E\rho^2$, two G-theory coefficients typically used to quantify reliability (Brennan 2001, Laenen *et al* 2008).

However similar, the two measures of reliability have different interpretations, and thus provide different information. The R_T coefficient is, regarding interpretation, closest to the coefficients based on the classical methods (e.g. Pearson correlation or intraclass correlation). In a longitudinal context it quantifies the average reliability over the repeated measurements, however, separate values per time point can also be obtained. The R_Λ coefficient, on the other hand, involves a new way of looking at reliability in a longitudinal context. This measure expresses the reliability of the entire sequence of observations. It captures the idea that every new measurement brings additional information about the patients, fully corresponding with the clinical believe that, the longer a patient is followed, the more reliable will be the conclusions about this patient. A logical result is then that the value of R_Λ increases with the number of measurements.

3.3 Data Analysis

We will estimate both reliability coefficients, R_T and R_Λ, for the three scales HAMD, MADRS, and HAMA, based on the clinical trial data introduced in Section 3.1. As explained in Section 3.2, the methodology rests upon the estimated variance-covariance parameters ($\hat{V}$ and $\hat{\Sigma}$), which result from fitting a linear mixed model. In order to obtain unbiased reliability estimates it is of utmost importance to find a statistical model that describes the data reasonably well. To this effect, model building guidelines, as laid out in, for example, Verbeke and Molenberghs (2000,

Ch. 9) ought to be followed.

For the data at hand an ample model building exercise was carried out. An elaborate mean model ($X_i\boldsymbol{\beta}$) was adopted, taking time as a categorical variable, and further including investigator, treatment, and treatment by time interaction. Such a general mean model minimizes the risk of bias in the estimates of the variance parameters (Diggle, Liang, and Zeger 1994) and allows to model, for example, quadratic trends that are typically observed in this type of data. To optimally model the random effects ($\boldsymbol{b_i}$) and the measurement error ($\boldsymbol{\varepsilon_i}$), many different potential models were considered and the data were used to select the best one. The Akaike's information criterion (AIC) was applied for selecting the best fitting model and parameter estimation was based on the so-called restricted maximum likelihood method (REML), a bias-reducing version of maximum likelihood (Verbeke & Molenberghs 2000).

Table 1 summarizes the covariance structure of the final models obtained for the three scales in each trial. Models selected for the first trial indicate a subject-specific linear evolution over time whereas the models for the second trial indicate a subject-specific quadratic evolution over time. Note that these evolutions indicate individual deviations from the average time evolutions. This obviously illustrates that the classical steady-state assumption is unrealistic in this scenario. Additionally, all models indicate the presence of correlated error terms captured by the 'heterogeneous autoregressive' and 'banded unstructured' variance-covariance matrix Σ. This finding clearly hints on a violation of the classical assumption of uncorrelated errors. Finally, all but one of the selected models indicate that the variability changes over time, once again showing a disagreement with the classical assumptions.

The previous analysis manifestly reveals the limitations of the classical methods to approach the reliability problem in a longitudinal framework. Most of the fundamen-

Table 1: *Selected models for the three scales.*

	Scale	Random effects structure	Structure of Σ
Trial 1	HAMD	linear slope	heterogeneous autoregressive
	MADRS	linear slope	heterogeneous autoregressive
	HAMA	linear slope	banded unstructured
Trial 2	HAMD	quadratic slope	heterogeneous autoregressive
	MADRS	quadratic slope	heterogeneous autoregressive
	HAMA	quadratic slope	autoregressive

tal assumptions needed in the classical setting are violated in the present studies. Carrying out the estimation of reliability in this scenario using classical methods like CTT or G-theory will almost certainly lead to biased estimates. Probably giving an overoptimistic appraisal of the reliability coefficients.

The lower half of Figure 1 shows the residual patient profiles for the three scales, resulting from the best fitting models in trial 1. These are the observed patient scores after substraction of the model predictions. If the models successfully capture the general features of the data generating mechanism, then these graphs should not show any systematic pattern over time. Clearly, no systematic pattern emerges from the graphs, indicating that the models capture the most important data features reasonably well. Further, Figure 2 plots the observed score values for three randomly chosen patients in trial 1 together with the scores predicted by the model. Here again, a reasonable agreement between the models and the data is observed, reinforcing our confidence in the results of the model building step. Similar results (not shown) were found for trial 2.

Eventually, we want to point at the fact that incomplete observations are found for some patients, as can be seen also in Figure 1. However, as indicated before, the model fitting has a likelihood basis, and therefore the methodology remains

Table 2: *Reliability results for the three scales.*

	Scale	R_T	CI_{R_T}	R_Λ	CI_{R_Λ}
Trial 1	HAMD	0.493	[0.405; 0.581]	0.829	[0.734; 0.895]
	MADRS	0.474	[0.378; 0.571]	0.812	[0.704; 0.886]
	HAMA	0.612	[0.545; 0.676]	0.955	[0.897; 0.980]
Trial 2	HAMD	0.629	[0.513; 0.731]	0.932	[0.872; 0.966]
	MADRS	0.692	[0.603; 0.769]	0.977	[0.957; 0.988]
	HAMA	0.675	[0.601; 0.741]	0.964	[0.930; 0.986]

statistically valid if the missing data mechanism is missing at random (Rubin 1976), in the sense that missingness is allowed to depend on observed data but, given these, not further on unobserved data.

Once sufficiently adequate models have been selected, reliability can be estimated using the variance components estimates emanating from these models. SAS macros to carry out all the necessary computations can be obtained from the authors.

4 Results

Reliability estimates are obtained separately for both clinical trials. The general results are presented in Table 2, with important details highlighted in Figure 3.

Let us first compare the HAMD and MADRS depression scales. The two graphs at the top of Figure 3 show the R_T values for both scales at each time point. These graphs illustrate that both scales perform rather poorly at the beginning of the trials. However, we can see that in both studies, the R_T values increase with time. For trial 1 we observe a gradual increase, whereas in trial 2 the increase is more abrupt. Arguably, such an increase could have been induced by a learning effect of the raters, stemming from gaining experience and/or enhanced familiarity with a patient during follow-up.

To compare the two scales, it is also useful to look at the general R_T values (Table 2) that give the average reliability over the different time points. First of all, let us note that moderate values of reliability were obtained for all the scales in both trials. These values are clearly lower than the ones frequently reported in the previous literature. Many reasons could explain this discrepancy. First, reliability is a population depending concept and therefore, different studies can in principle lead to different results. Second, the setting studied here significantly differs from the ones considered in classical reliability studies. Therefore, extrapolating the results obtained in these classical studies to more complex scenarios, like the one encountered here, can be misleading. This, once again, emphasizes the importance of studying reliability within a longitudinal setting and using longitudinal data.

Interestingly, regarding the point estimates in the first trial, HAMD performs slightly better than MADRS, whereas in trial 2 the opposite behavior is observed. Irrespective of these small differences in the point estimates, Table 2 reveals that the confidence intervals for R_T of the two scales largely overlap in both trials. Clearly, based on the present data, we encounter no evidence that MADRS is a more reliable scale than HAMD. This finding is somehow unexpected, taking into account that MADRS was created to address some of the limitations of HAMD.

Further, note that the reliability estimates for the two scales are clearly higher in the second trial than in the first one. Reliability is known to be a population-dependent concept, and will generally be estimated higher in more heterogenous groups. However, it is highly unlikely that this can explain the observed difference between the two trials since both studies were developed from one protocol and they were identical in every way. Other factors might have had an influence as well, such as training, experience, and quality of the raters. Also on this matter, equality of the two trials was aimed for. At a single start up meeting, all sites in both studies

were present to be trained on the protocol and to qualify raters. Investigative sites were randomly selected to be part of either trial. But there is no guarantee that this random assignment truly equalized quality of sites and raters.

Even though it is difficult to identify the reasons for the differences in reliability between the two trials, it is very interesting to relate this finding to the clinical outcomes of the studies. Both studies tested 3 arms of what are now proven to be effective doses of anti-depressants. Trial 1, however, had worse separation from placebo than trial 2 (Mallinckrodt *et al* 2003). The finding that the reliability of the measurements was also lower in the first trial might explain why the clinical effects were stronger in the second trial. This finding illustrates that measurement error or low reliability can have an effect on the results found in clinical studies, as emphasized by Fleiss (1987) and Lachin (2004).

Let us now turn to the second reliability measure, R_Λ, quantifying the reliability of the *accumulated* observations. When we measure the patients once, we obtain a certain amount of information. By measuring a second time, we can only increase the amount of information on the patient even if it comes contaminated by measurement error. This fact is nicely captured by the R_Λ. The lower half of Figure 3 shows the cumulative R_Λ values, over the different time points. At the first time point, R_Λ expresses the reliability of the first measurement, which is equal to R_T at the first time point. At the second time point, R_Λ captures the reliability of the information contained in the first and the second measurement combined, and so on. The values shown in Table 2 present the results for the entire study, expressing the situation at the last time point. R_Λ illustrates that, whenever a scale has low reliability, reliable results can still be obtained when the scale is applied repeatedly over time and the repeated outcomes are considered together. Of course, the lower the reliability of the scale at each time point, the more measurements will be needed to obtain a

pre-specified degree of cumulative reliability. The practical implications of these findings are considerable. Most of the rating scales used in psychiatry, and in many other areas, have a strong subjective component. It is therefore expectable that only moderate values of reliability are going to be achieved for several of these instruments. However, the R_Λ illustrates that reliable results can be obtained even with quite "imperfect" instruments as far as a sufficient number of assessment are carried out. It also gives us the possibility of estimating the number of observations needed to achieve a pre-specified level of reliability with a given scale in a given population. For instance, Figure 3 shows that, in the first trial, a value of 0.80 was reached only at the last measurement. In the second trial, 5 and 4 measurements, respectively, were needed to reach the same level of reliability for both HAMD and MADRS.

While in the first trial, the cumulative evolutions of R_Λ are very similar for both depression scales, a better performance is observed for MADRS compared to HAMD at the beginning of the second trial. The relatively high reliability for MADRS at the first time point gives this scale a head start. Towards the end of the trial, HAMD has caught up with MADRS, leading to a small difference in the final R_Λ values, as shown in Table 2.

To find out whether, in the second trial, the R_Λs for MADRS and HAMD differ significantly at the beginning of the study, we plot the 95% confidence bands for the cumulative R_Λ values for both scales, as shown in Figure 4. The figure shows wide confidence intervals for the earlier time points, while they get narrower towards the end of the study, when more information becomes available. The intervals for the two scales overlap at any of the time points, thereby failing to show evidence of MADRS performing better than HAMD.

Let us now look at the results for HAMA. This particular scale measures anxiety and should therefore not be compared directly to the two depression scales. Table 2 shows somewhat better reliabilities in the second trial, which is in agreement with earlier findings. However, the differences are not too large. The average reliabilities, R_T, are 0.61 and 0.68, respectively, indicating a decent, however not excellent, reliability. The results for trial 1 clearly illustrate that, even when the R_T values are stable over time, the total information, as expressed by R_Λ, still increases. When a level of 0.80 is aimed at, four measurements are needed in case of the first trial and three in case of the second trial.

We further analyze the differences in reliability estimates between different clinical populations. A dichotomization of the CGI Severity divided the group of patients into a group of less severely depressed patients (from "not ill" to "moderate") and more severely depressed patients (from "marked" to "extremely severe"), based on measurements taken two weeks before baseline. Table 3 shows that the group of less severely depressed patients gives rise to lower reliability estimates in all the cases, when compared to more severely depressed patients. This may hint on a better capacity of the scale to differentiate between severely depressed patients than between less severely depressed patients. However, when we look at the 95% confidence intervals, we see that for any of the comparisons there is partial overlap between the two intervals. The results thus should be interpreted with extreme care. Interesting though, is to know that in both trials almost two thirds of the patients suffer from mild depression at the beginning of the study. This might partly explain the relatively low reliability estimates that were obtained in the two trials.

Further, we distinguish between non-responders and responders. A patient is considered as a responder in case of 50% change from baseline at the endpoint visit. Note that the latter is based on HAMD measurements. Table 4 shows that higher

Table 3: *R_T estimates and 95% confidence intervals for less and more severely depressed patients.*

	Scale	less severely depressed		more severely depressed	
Trial 1	HAMD	0.453	[0.356; 0.554]	0.579	[0.466; 0.686]
	MADRS	0.448	[0.347; 0.555]	0.535	[0.410; 0.656]
	HAMA	0.602	[0.526; 0.674]	0.632	[0.541; 0.715]
Trial 2	HAMD	0.599	[0.469; 0.716]	0.674	[0.552; 0.776]
	MADRS	0.667	[0.568; 0.753]	0.735	[0.638; 0.813]
	HAMA	0.649	[0.564; 0.726]	0.707	[0.621; 0.780]

Table 4: *R_T estimates and 95% confidence intervals for non-responders and responders.*

	Scale	non-responders		responders	
Trial 1	HAMD	0.523	[0.387; 0.656]	0.492	[0.366; 0.619]
	MADRS	0.528	[0.398; 0.653]	0.452	[0.324; 0.586]
	HAMA	0.625	[0.535; 0.707]	0.608	[0.522; 0.687]
Trial 2	HAMD	0.765	[0.669; 0.839]	0.497	[0.355; 0.639]
	MADRS	0.827	[0.758; 0.880]	0.573	[0.455; 0.683]
	HAMA	0.713	[0.614; 0.795]	0.655	[0.568; 0.733]

reliability estimates are found for non-responders compared to responders. In two of the six comparisons the 95% confidence intervals are completely separated, indicating significant differences. This is true for the HAMD and MADRS in the second trial. These results are consistent with the ones previously found. Indeed, here again the scales seem to be more capable to differentiate between subjects in the population defined by the "problematic" cases, i.e., patient with no response.

5 Discussion

We will recapitulate the key messages of the paper based on the aims that were outlined in Section 2.

Obtaining unbiased reliability estimates for HAMD, MADRS and HAMA

Despite numerous psychometric flaws of the HAMD (Zimmerman *et al* 2005), the inter-rater and test-retest reliabilities reported in the literature are mostly good. Bagby *et al* (2004) analyzed 70 studies and reported a range between 0.81 and 0.98 for both reliability types, based on Pearson correlation coefficients. In the present analysis, we obtained average reliabilities, based on the R_T coefficient, around 0.50 and 0.60 respectively, for two different studies. The fact that these numbers are lower than the reliabilities mentioned in the literature can have different reasons. As indicated before, reliability is a population-dependent concept and tends to be lower in more homogeneous populations. The studies on which the present estimates are based were conducted in a patient segment suffering from a major depressive disorder, likely reducing variability between the patients. It is not always clear on which populations the reliability estimates in the literature are based. A second plausible reason for the observed difference might be due to the analysis method. Studies mentioned in the literature mostly use the Pearson correlation coefficient as reliability measure. If however, in a test-retest design, patients have evolved during the time period between the two measurements, or, if a memory-effect is into play, the correlation coefficient overestimates the reliability. The methods used in this paper allow to correct for such features and provide unbiased reliability estimates for the HAMD. The same is, of course, true for the MADRS and the HAMA.

Optimizing the use of the rating scales

In this paper we applied two different measures for reliability. The R_T coefficient

is most closely related to the classical measures. Values obtained for the R_T can easily be compared to values obtained by classical methods, such as the Pearson or intraclass correlation. The second coefficient, R_Λ, involves a relatively new way of thinking about reliability in the context of repeated measurements. This coefficient captures the reliability of the entire sequence of measurements. Increasing the number of time points then leads to an increase of the total information and therefore an increase of reliability. This implies that, even if the only available rating scale has a rather low reliability in a single administration, the same scale can still provide reliable information when it is applied repeatedly. Applying this concept to the case study data, we found that between 3 and 4 measurements were needed to achieve a reliability of 80% for all the scales in the second clinical trial. In the first trial more measurements were needed to obtain the same level of reliability. However, at the end of both trials, i.e. after 6 measurements, all scales arrived above this level of reliability. This lets us conclude that a scheme of 6 evaluations per patient should be enough to obtain reliable results with any of these instruments in the population of patients suffering from a major depressive disorder.

Comparing the reliability of HAMD and MADRS in a longitudinal setting

The many psychometric problems reported on the HAMD (Zimmerman *et al* 2005, Bagby *et al* 2004) has led to the concern that flawed outcome measures might hide treatment effects of the newer generation of antidepressant medications. This, in turn, has led to the conclusion that the HAMD should be replaced by an alternative scale that solves these problems. One scale proposed to this end has been the MADRS. Because in both clinical trials the two rating scales were applied, we had the chance of comparing these scales with respect to their reliabilities. We did not find any evidence, however, of one scale outperforming the other. Our study thus confirms the findings of Maier *et al* (1988) on inter-rater reliabilities.

Studying reliability in different clinical populations

The data indicate that the HAMD as well as MADRS are more reliable in a population of non-responders than in a population of responders. For the HAMA the results are less obvious; in the first trial reliability estimates are practically identical. Furthermore the data suggest that all three scales are somewhat more reliable in a group of more severely depressed patients compared to less severely depressed patients. However, this conclusion needs to be taken with care.

Illustrating the advantages of more advanced analysis methods for reliability

The methods for reliability estimation used in this paper are clearly much more complex than the usual methods based on classical test theory. Nevertheless, they imply important advantages. In the first point we have mentioned the advantages of the linear mixed models to obtain unbiased reliability estimates in situations where the classical approach might lead to an overestimation. A direct consequence of the fact that linear mixed models can adopt and therefore correct for several disturbing factors is that reliability estimates can be derived from the longitudinal clinical trial data. This in contrast to the usual practice of organizing an additional reliability study on a subgroup of the sample to evaluate a scale's reliability in a new population. Apart from this cost-saving option, this approach will also lead to a gain in precision. When the clinical data can be used for reliability estimation, it follows that all the patients are involved in the reliability analysis. Reliability estimates will logically be estimated with a much larger precision, compared to estimations based on subsamples of 20 or 30 patients. Finally, this new methodology allows to avoid dangerous extrapolations from studies with a different design and based on restrictive assumptions.

Furthermore the R_Λ coefficient might be very useful in evaluating past trials as well

as in planning new trials. The power to detect treatment effects increases with sample size as well as with reliability (Fleiss 1986). We have seen that the number of repeated measurements per patient increases the reliability of the accumulated measurements within a trial. When estimates of R_Λ are known for a scale, a trade off can be made between the sample size on one hand and the number of repeated measurements on the other hand.

6 References

Bagby, R.M., Ryder, A.G., Schuller, D.R., & Marshall M.B. (2004). The Hamilton Depression Rating Scale: Has the gold standard become a lead weight? *American Journal of Psychiatry*, **161**, 2163–2177.

Bost, J.E. (1995). The effect of correlated errors on generalizability and dependability coefficients. *Applied Psychological Measurement*, **19** (2), 191–203.

Brennan R.L. (2001). Generalizability Theory. New York: Springer-Verlag.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles.* New York: John Wiley.

Cronbach, L.J., Rajaratnam, N., & Gleser, G.C. (1963). Theory of generalizability: a liberation of reliability theory. *British Journal of Statistical Psychology*, **16**, 137–163.

DeShon, R.P., Ployhart, E. & Sacco, J.M. (1998). The estimation of reliability in longitudinal models. *International Journal of Behavioural Development*, **22**, 493–515.

Diggle, P.J., Liang, K.-Y., Zeger S.L. (1994). *Analysis of Longitudinal Data.* Oxford Science Publications. Oxford: Clarendon Press.

Fleiss, J.L. (1986). *Design and Analysis of Clinical Experiments.* New York: John Wiley & Sons.

Gueorguieva, R., & Krystal, J.H. (2004). Move over ANOVA: Progress in analyzing repeated measures data and its reflection in papers published in the Archives of General Psychiatry. *Arch Gen Psychiatry* **61**, 310–317.

Hamilton, M. (1959). The assessment of anxiety states by rating. *Br J Med Psychol*, **32**, 50–55.

Hamilton, M. (1960). A rating scale for depression. *J Neurol Neurosurg Psychiatry*, **23**, 56–62.

Lachin, J.M. (2004). The role of measurement reliability in clinical trials. *Clinical Trials*, **1**, 553–566.

Laenen, A., Alonso, A., and Molenberghs, G. (2007). A measure for the reliability of a rating scale based on longitudinal clinical trial data. *Psychometrika*, **73**, 443–448.

Laenen, A., Alonso, A., Molenberghs, G., & Vangeneugden, T. (2008). Reliability of a longitudinal sequence of scale ratings. *Psychometrika* (Accepted for publication).

Laird, N.M., & Ware, J.H. (1982). Random effects models for longitudinal data. *Biometrics*, **38**, 963–974.

Leon, A.C., Mallinckrodt, C.H., Chuang-Stein, C., Archibald, D.G., Archer, G.E., & Chartier, K. (2006). Attrition in Randomized Controlled Clinical Trials: Methodological Issues in Psychopharmacology. *Biological Psychiatry* **59**, 1001–1005.

Lord, F.M., & Novick, M.R. (1968). *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley.

Maier, W., Philipp, M., Heuser, A., Schlegel, S., Buller, R., & Wetzel, H. (1988). Improving depression severity assessment: Reliability, internal validity and

sensitivity to change of three observer depression scales. *Journal of Psychiatric Research*, **22**, 3–12.

Mallinckrodt, C.H., Goldstein, D.J.,Detke, M.J. Lu, Y., Watkin, J.G., & V.Tran, P. (2003). Duloxetine: a new treatment for the emotional and physical symptoms of depression. *Primary Care Companion. Journal of Clinical Psychiatry*, **5**, 19–28.

Mallinckrodt, C.H., Watkin, J.G., Molenberghs, G., & Carroll, R.J. (2004). Choice of The Primary Analysis in Longitudinal Clinical Trials. *Pharmaceutical Statistics* **3**, 161–169.

Molenberghs, G., & Kenward, M.G. (2007). *Missing Data in Clinical Studies.* Chichester: John Wiley & Sons.

Montgomery, S.A., & Åsberg, M. (1979). A new depression scale designed to be sensitive to change. *British Journal of Psychiatry*, **168**, 594–597.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, **63**, 581–592.

Smith, P.L., and Luecht, R.M. (1992). Correlated effects in generalizability studies. *Applied Psychological Measurement*, **16** (3), 229–235.

Vangeneugden, T., Laenen, A., Geys, H., Renard, D. & Molenberghs G. (2004). Applying linear mixed models to estimate reliability in clinical trial data with repeated measurements. *Controlled Clinical Trials*, **25**, 13–30.

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

Zimmerman, M., Posternak, A., & Chelminski I. (2005). Is it time to replace the Hamilton depression rating scale as the primary outcome measure in treatment

studies of depression. *Journal of Clinical Psychopharmacology*, **25**,105–110.

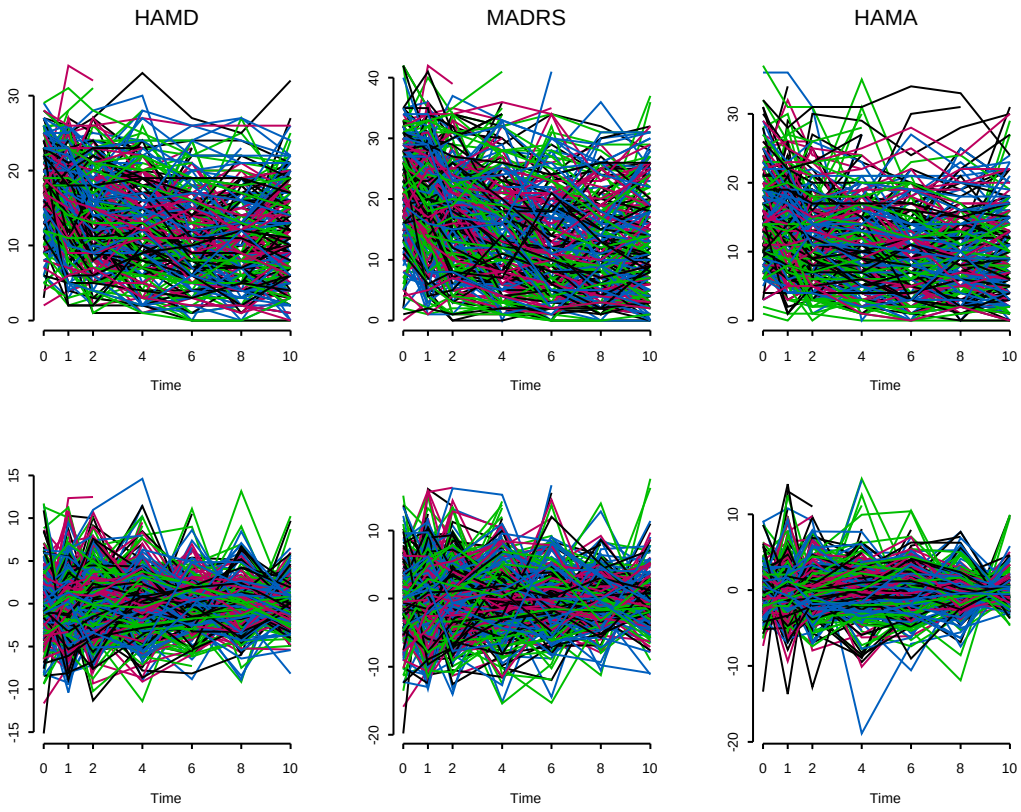

Figure 1: *Individual patient profiles for three rating scales: observed (top) and residual (bottom) profiles.*

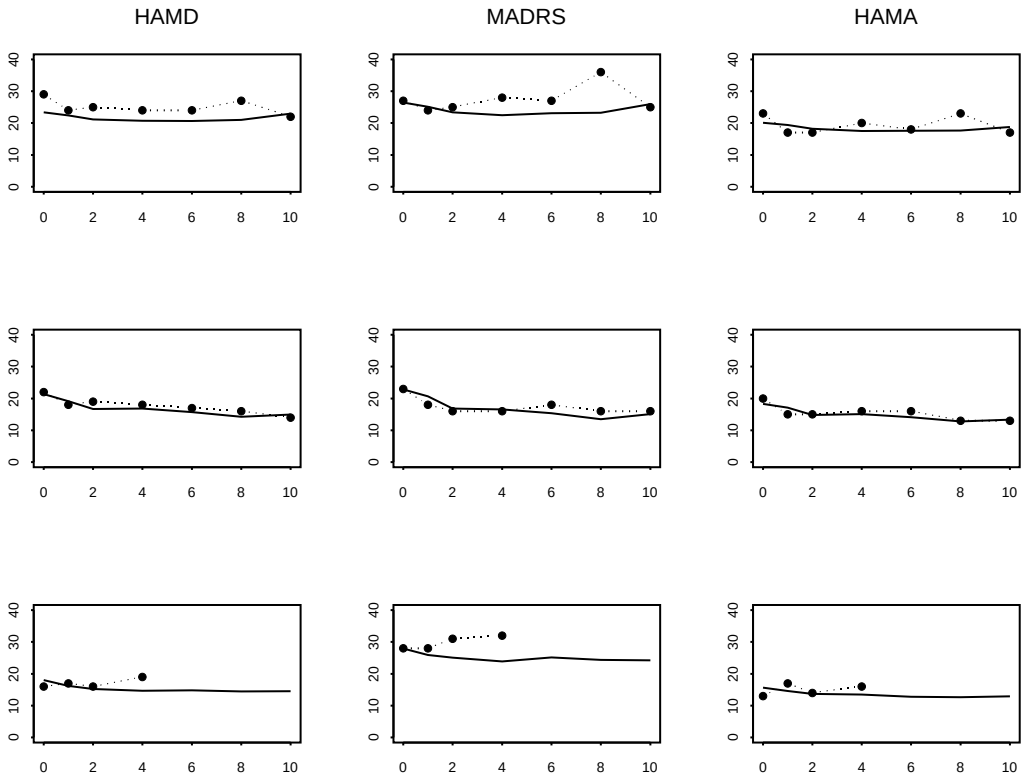

Figure 2: *Individual observed profiles (dots) and fitted profile (solid line) for three randomly selected patients.*

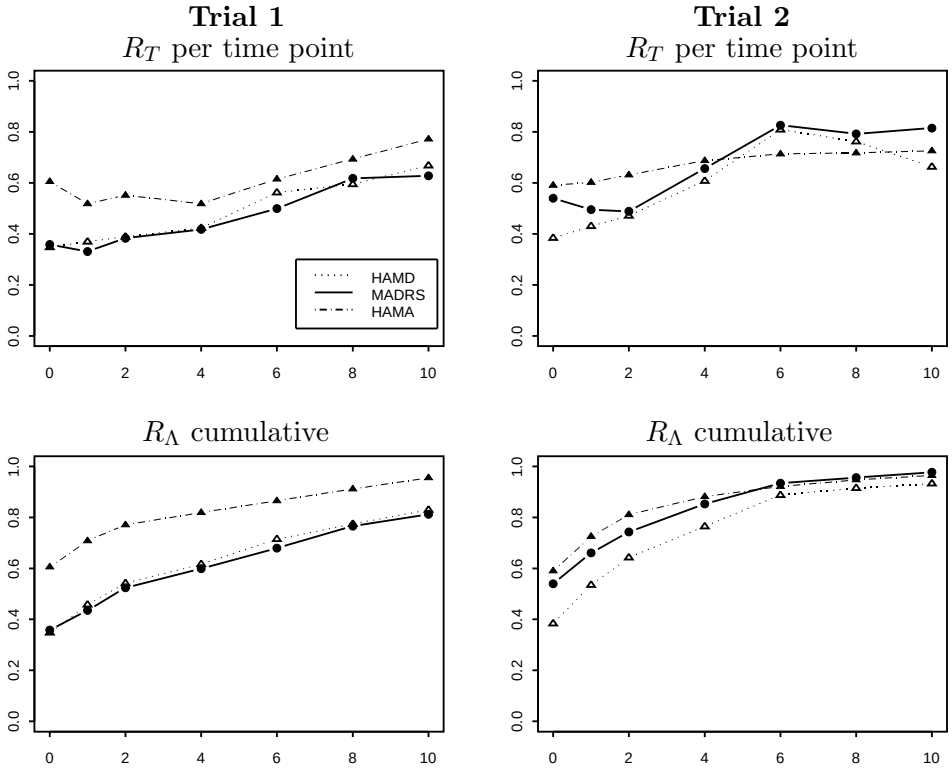

Figure 3: *R_T per time point and R_Λ cumulative over time points.*

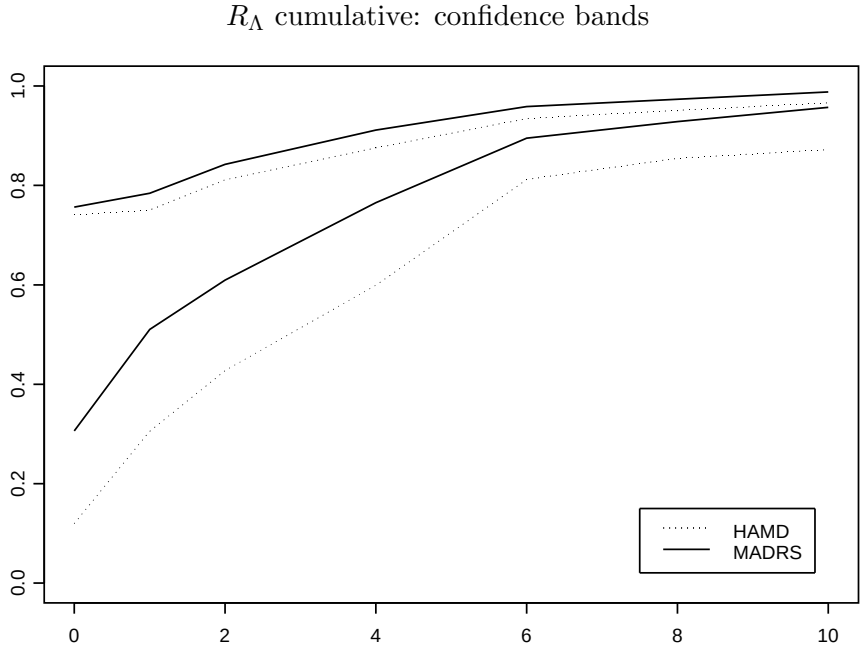

Figure 4: *Trial 2. 95% confidence bands around R_Λ cumulative over time points.*