To fail or not to fail : clinical trials in depression

Sante, G.W.E.

Chapter

4

Comparative analysis of the sensitivity of the individual items of the Montgomery Asberg depression rating scale to response and its consequences for the assessment of efficacy

Gijs Santen, Meindert Danhof, Oscar Della Pasqua

Submitted for publication

ABSTRACT

The most frequently used endpoints in depression trials are the Hamilton depression rating scale (HAMD) and the Montgomery Asberg depression rating scale (MADRS). An increasing body of evidence is available which suggests that the HAMD is not a sensitive measure of treatment effect. In fact, subscales have been shown to consistently perform better than the full HAMD scale. In the current investigation, we explore the sensitivity of the individual items of the MADRS and compare the consequences of selecting the HAMD, its subscales or the MADRS as primary endpoint in the analysis of efficacy in depression studies.

For this analysis, data from two double-blind, randomised, placebo-controlled, clinical studies were used in which the HAMD and MADRS were measured concurrently as efficacy endpoint. A graphical approach was applied for the evaluation of the sensitivity of individual items to response, as defined by a ≥50% decrease in HAMD relative to baseline values. Based on a bootstrap technique and the mixed model for repeated measures (MMRM), we illustrate the impact of differences in the sensitivity of the primary endpoint on the detection of statistical differences in treatment effect.

In contrast to the HAMD, our item-by-item analysis of the MADRS reveals that all individual items are sensitive to response, irrespective of treatment type. However, whilst the MADRS was consistently more sensitive to response than the HAMD, some of the subscales of the HAMD outperformed the MADRS in the detection of treatment effect.

In conclusion, the use of MADRS is recommended as primary endpoint in efficacy trials with anti-depressant drugs when regulatory constraints prevent the use of subscales as endpoint.

INTRODUCTION

There are at least two important reasons to investigate the sensitivity of endpoints used in clinical trials in depression. Firstly, more than 50% of the performed trials fail, even if efficacious doses of known antidepressants are used (Khan *et al.*, 2002). Secondly, the use of more sensitive endpoints may yield an increased effect size and facilitate the detection of significant treatment effect (chapter 3). Hence, fewer patients need to be enrolled and study duration and costs will decrease.

Although many factors may explain failure of depression trials, such as inadequate sample size, sub-optimal doses, inefficacious drugs and inadequate duration of the trial, we believe that endpoint sensitivity is a major contributor to the problem and one that can be readily investigated using historical data.

In contrast to other therapeutic areas for which objective diagnostic criteria are available, the use of rating scales in psychiatric diseases, such as the Hamilton depression rating scale (HAMD) (Hamilton, 1960) and Montgomery Asberg depression rating scale (MADRS) (Montgomery and Asberg, 1979), has evolved from an empirical assessment of clinical symptoms and remains uncontested in clinical practice. In fact, they are not true diagnostic instruments, but are methods of comprehensively surveying the type and magnitude of symptom burden present, and are therefore considered to be measures of illness severity. The HAMD and MADRS are each conducted as semi-structured observer-rated interviews (table 1). However, the magnitude of item scaling differs between the two instruments. The MADRS has a fixed scaling of seven points (from 0 through 6), while the scoring on the HAMD ranges across a smaller number of scaling points, and varies from item to item.

Since its introduction in 1960, the HAMD has been the most widely used endpoint in antidepressant trials. Criticism on this scale, particularly on its multidimensionality and unsuitability to monitor changes upon treatment (Bech and Rafaelsen, 1980; Moller, 2001), has led to the evaluation and development of new scales. Broadly, they can be divided into two categories. On the one hand, subscales of the HAMD were devised aggregating between 5 and 7 items (Bech and Rafaelsen, 1980; Maier and Philipp, 1985), which were shown to be unidimensional and more sensitive to treatment effect (O'Sullivan *et al.*, 1997; Faries *et al.*, 2000). On the other hand, completely new scales were created with the specific goal to be used to detect changes upon treatment (Bech and Rafaelsen, 1980; Montgomery and Asberg, 1979). The most important scale in this respect is the MADRS, which was introduced in 1979. Since then, many studies have used the MADRS as primary endpoint in anti-depressant trials. To our knowledge, the HAMD, its subscales and the MADRS have not yet been compared previously with respect to their sensitivity to detect treatment effect.

In a recent publication (chapter 3) we have shown that not all items of the HAMD are equally sensitive to response, which was defined as a reduction of at least 50% from baseline values, irrespective of treatment type (placebo or active drug). This methodology allowed us to derive a new subscale (HAM-D_7), retaining seven items which show a clear time-dependent pattern separating responders from non-responders within the patient population. Subsequently, we have used the mixed model for repeated measures (MMRM) (Mallinckrodt *et al.*, 2001) and a bootstrapping technique to demonstrate the impact of this subscale on the estimation of the significance level of treatment effect and corresponding statistical power, as compared to the full HAM-D_{17} scale. These findings provide further evidence for the need to reconsider clinical trial practice, allowing for fewer patients to be enrolled in the evaluation of experimental drugs. In addition to the reduction of false negative results, the introduction of alternative scales bears an important ethical aspect in that one can ensure fewer patients are exposed to placebo treatment.

Given that the MADRS was especially designed to detect treatment effect, the application of the same methodology to the MADRS is an interesting prospect. The aim of the current investigation was therefore to evaluate the sensitivity of individual items of the MADRS to response (irrespective of treatment type), followed by a comparison of the estimates of treatment effect size obtained by the use of the MADRS, HAMD and its subscales as efficacy measure in clinical studies in depression.

METHODS

Study data

Data from two studies in major depressive disorder (MDD) were obtained from GlaxoSmithKline's clinical database. To meet the objectives of the current investigation, study selection was based on the availability of concurrent assessments of the HAMD and MADRS, frequency of clinical visits, total duration of the trial, as well as well-defined criteria regarding patient population, design and dosing regimen. Patients should be diagnosed with major depressive disorder and abstain from any other concomitant antidepressant medication. Studies should be randomised, double blind and placebo-controlled, with treatment allocation including different dose levels and titration schedules.

In study 1 four fixed doses of paroxetine (10, 20, 30 and 40 mg) were investigated (Dunner and Dunbar, 1992). In this study, 50 patients were enrolled in the placebo arm and 100 patients in each treatment arm. The HAM-D$_{17}$ (Hamilton, 1967) and MADRS were assessed at baseline and weeks 1, 2, 3, 4, 6, 9 and 12 after start of treatment. The data of this study was also included in the evaluation of HAMD subscales in chapter 3.

Study 2 was performed according to a dose-escalation design in which paroxetine (10-50 mg/day) was compared to imipramine (65-275 mg/day) (Feighner *et al.*, 1993). A total of 717 patients were evenly distributed among the treatment arms. The HAM-D$_{17}$ (Hamilton, 1967) and MADRS were assessed at weeks 1, 2, 3, 4 and 6 after start of treatment. Further details on the patient population and the study design are available in the original publications of the study results (Dunner and Dunbar, 1992; Feighner *et al.*, 1993).

In addition to the requirements for study design, study population and comparable clinical assessments, it is important to rule out the influence of concomitant medication and dropout on the accuracy of the proposed analysis. There were no adverse events or other non-specific factors leading to a dropout rate different from what is commonly observed in depression trials. As per protocol, hypnotics or psychotropic medication was not allowed during treatment.

Sensitivity analysis

In order to assess the sensitivity of each item to clinical response, the study population was split in a responder and non-responder subset. Patients were considered responders if their HAM-D$_{17}$ was reduced at least 50% from baseline at any time during the trial. Even though a definition of response based on the MADRS could have been applied, we have chosen to use the HAM-D$_{17}$ as the gold standard throughout this investigation to allow consistent comparison between scales.

Dichotomisation and pooling of the data from different treatment groups was performed after a preliminary evaluation showed no differences in the time course of response between placebo and active treatment, or any disparity in the time course of the MADRS items across active treatment groups in responders and non-responders. As defined in the study protocols, each patient was observed on six to eight occasions. The observations were grouped by week of visit. Observations in week 9 and 12 were only made in study 1. The time course of response was analysed by showing the proportion of patients scored with each value for the individual item (Jonsson, 2004). This procedure enables extraction of information on three different levels. First, it provides evidence of the time-dependence for the onset and maintenance of response for each item separately. Second, it shows the discriminatory power of each specific item to separate responders from non-responders. Third, it reveals the specificity of each item to distinguish placebo response from drug response, by subsequent clustering of responders by treatment arm.

Following the aforementioned data clustering, graphical analysis based on pattern detection throughout the course of treatment was used to evaluate the sensitivity of individual items and potentially identify subscales that can better describe treatment effect. All graphical analyses were performed in the language and environment for statistical computing R (R Development Core Team, 2007). MADRS items were scored to be insensitive, slightly sensitive or sensitive to response, by examining the differences in the time course of responder and non-responder population. Sensitivity in this context was defined as the capacity of an item to distinctly vary with time (visit) and population type (responder *versus* non-responder). Sensitive items were therefore those items showing an unambiguous pattern by time and population type. In contrast, slightly sensitive and insensitive items were those showing modest changes or minor variation, respectively. The MADRS items were subsequently compared with the HAMD scale to explore the differences and phenomenology of the symptoms domains showing sensitivity to response.

Statistical power and population size

A statistical evaluation of the relevance of endpoint sensitivity in clinical trial design was performed by estimating the study power associated with each rating scale. A linear mixed-effects modelling approach for repeated measures (Mallinckrodt *et al.*, 2004) (MMRM) was used to evaluate the treatment effect in both studies, using the MADRS, the total HAM-D$_{17}$, and the published Bech and Rafaelsen (1980), Maier and Philipp (1985) and Santen *et al.* (2008) (chapter 3) subscales as endpoints in the analysis. The method was implemented in *proc mixed* in SAS (v9.1 for Windows, SAS Institute, Cary, NC, USA) on absolute change from baseline data. Baseline measurement, week and treatment were included as fixed effects, as were the treatment-week and baseline-week interactions. The random effects were specified using the */repeated* statement to account for serial within-subject correlation. A significance level of α=0.05 was used to determine the significance of treatment effect, which was determined as an average over all weeks.

To investigate the possibility of reducing the population size, bootstrapping was performed in SAS, sampling 1,000 new populations with a size between 50 and 150 patients from the existing studies. The replicated data sets were subsequently reanalysed using the MMRM, and the percentage of trials in which a statistically significant drug effect was found is reported. Although uneven randomisation occurred in study 1, equal group sizes were simulated in the bootstrap analysis.

RESULTS

Sensitivity analysis

All but one of the individual items of the MADRS were found to be sensitive to response. Given the degree of sensitivity of all MADRS items, an evaluation of a subscale of the MADRS was deemed irrelevant. Figure 1 shows that each item has a time dependent profile that easily distinguishes between the responder and non-responder populations. The only exception is item 5 (*reduced appetite*), which showed the least difference between responders and non-responders.

The differences in time course between the items of the MADRS and the corresponding items of the HAMD were investigated for responders and non-responders (figure 2). The items are shown by the symptom domain as summarised in table 1.

Given that the number of items in each symptom domain differs across scales, assumptions were made about their clinical equivalence. *Apparent sadness* and *reported sadness* were both linked to the HAMD-item *depressed mood*. Since no matching item exists for *inability to feel*, this item is not included in the comparison. *Inner tension* was closely matched by *psychic anxiety* and *reduced appetite* by *loss of appetite*. As loss of concentration is part of the HAMD-item *retardation*, the MADRS-item *loss of concentration* was linked to it. Finally, *pessimistic thoughts* was assumed to be equivalent to the

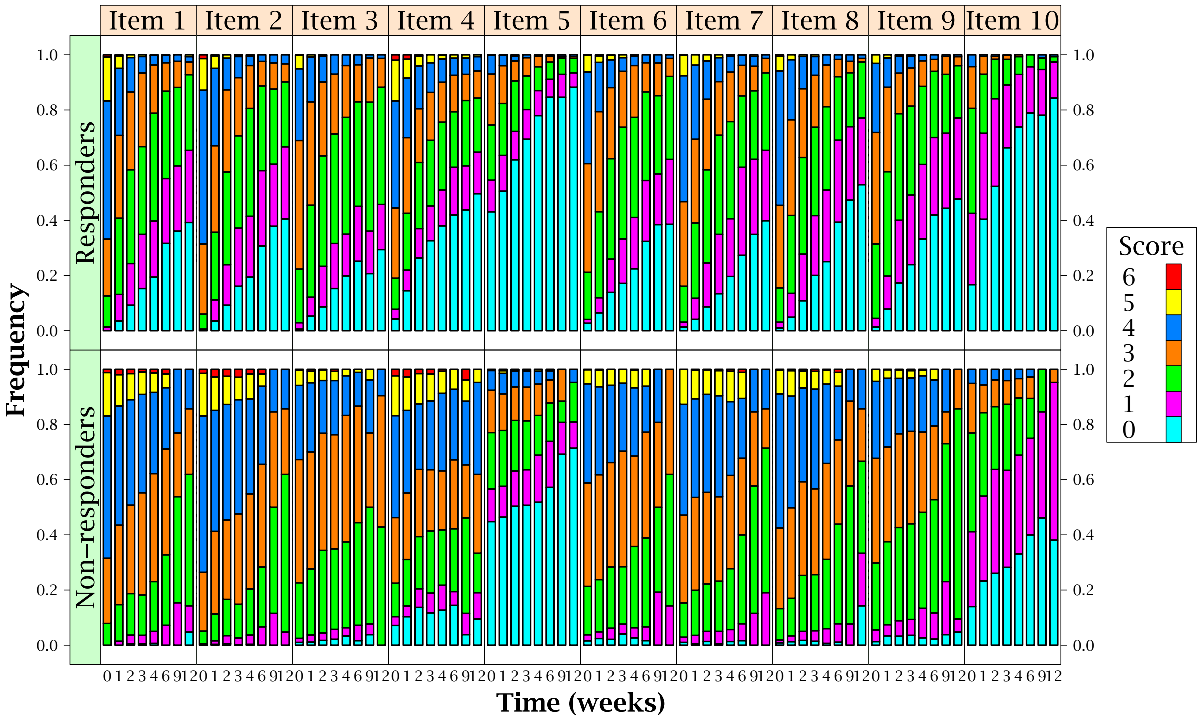
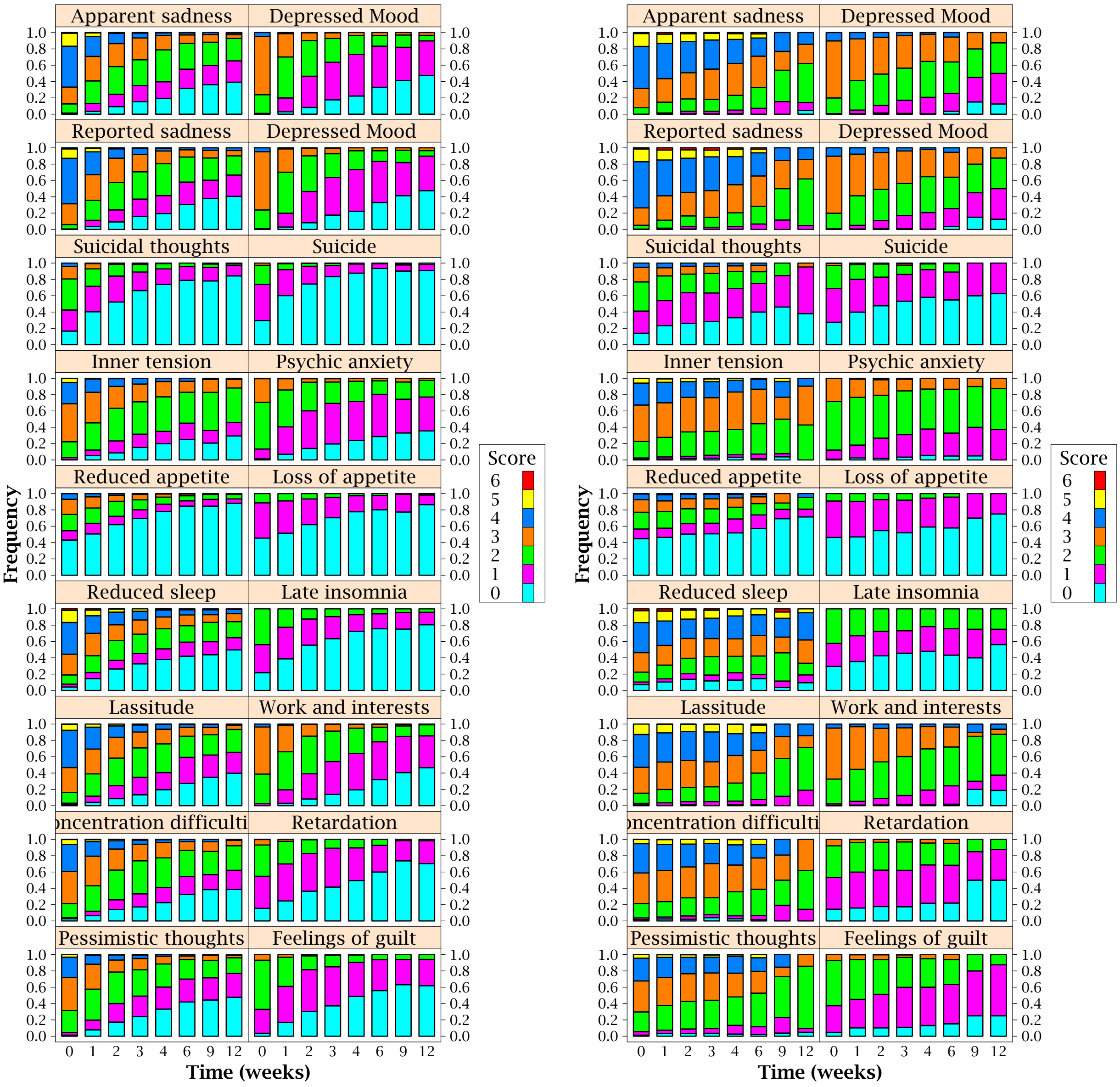


Figure 1. Time course and score distribution for items from the MADRS. Upper panel shows patterns in responders. Lower panel displays patterns in non-responders. The numbers correspond to the following items: 1=*apparent sadness*, 2=*reported sadness*, 3=*inner tension*, 4=*reduced sleep*, 5=*reduced appetite*, 6=*concentration difficulties*, 7=*lassitude*, 8=*inability to feel*, 9=*pessimistic thoughts*, 10=*suicidal thoughts*

HAMD-item *feelings of guilt*. From figure 2, it becomes evident that the time course of all items in each symptom domain is very similar, especially if some of the scores for the MADRS-items are grouped together.

For comparison purposes, figure 3 shows the results of the sensitivity analysis for the HAMD with the data from studies 1 & 2. Relative to the MADRS, there is clearly a much greater variability in the HAMD items with respect to the sensitivity towards response (difference between the upper and lower panels). Items such as *loss of insight* and *loss of weight* do not discriminate responders from non-responders at all.



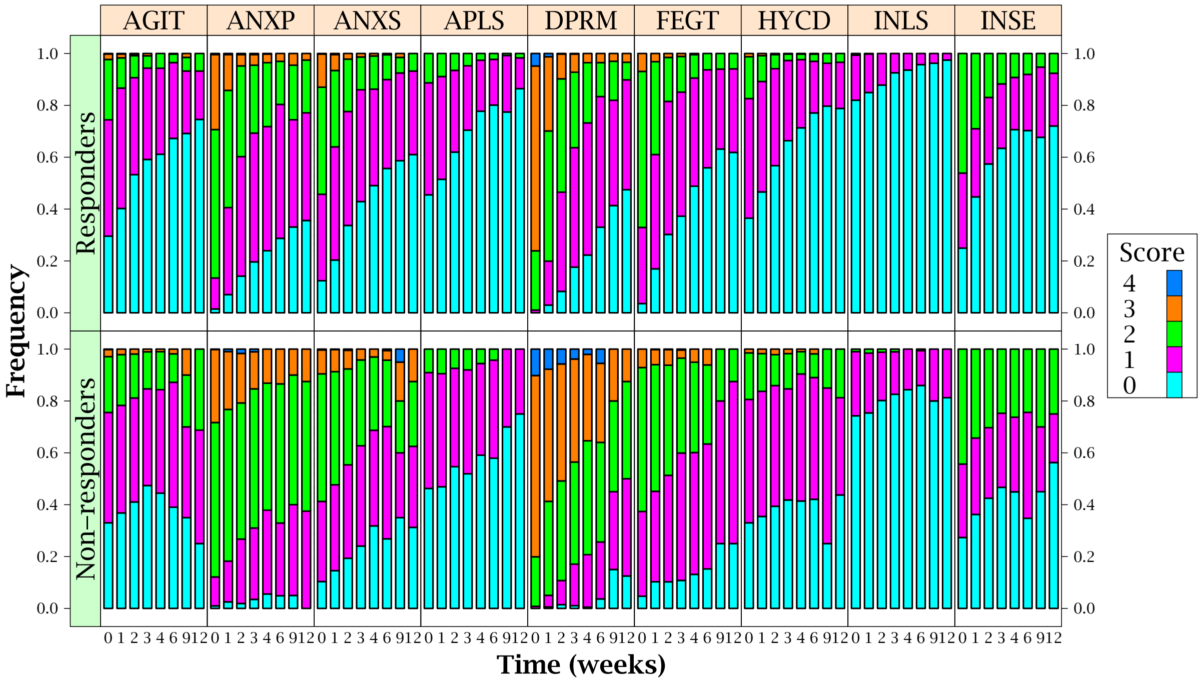
(a) Responders

(b) Non-responders

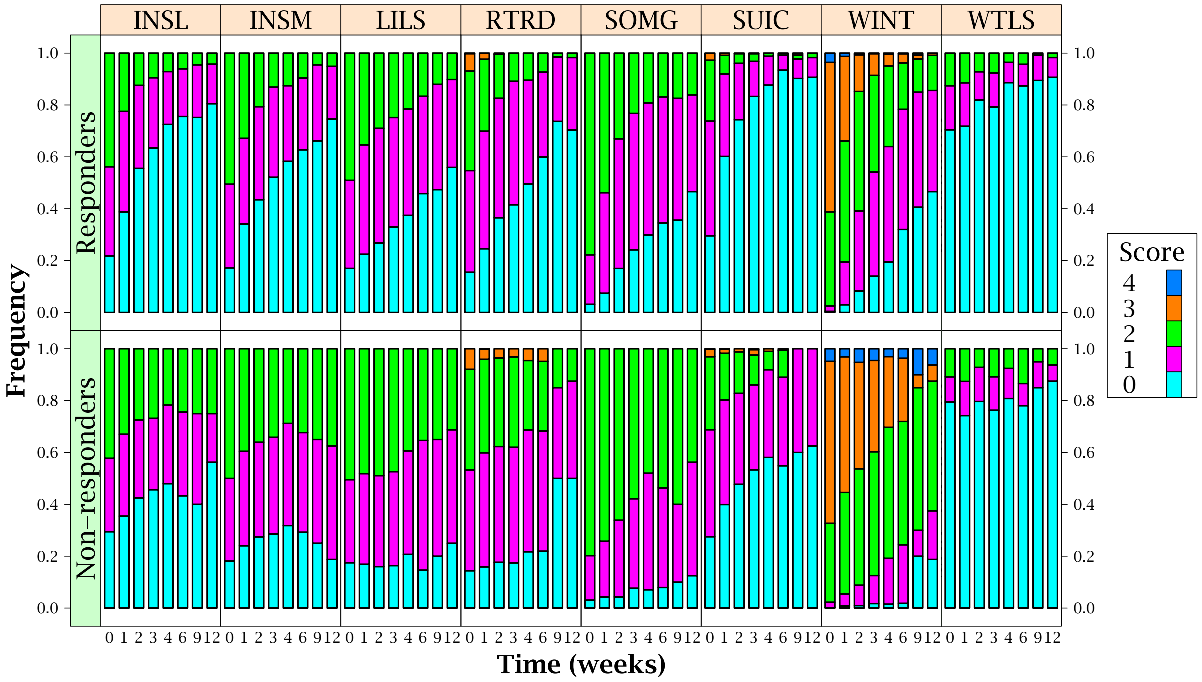
Figure 2. Time course and score distribution for items from the MADRS and corresponding items from the HAMD for (a) responders and (b) non-responders

Table 1. A list of the items of the MADRS and corresponding HAMD items by symptom domain. The HAMD items included in the subscales 1 and 2 (chapter 3) and the Bech HAM-D$_6$ are marked by an 'X'. Adapted from Fitzgerald (2007)

Symptom domain	MADRS	HAM-D$_{17}$	Subscale 1	Subscale 2	Bech HAM-D$_6$
Mood	Reported sadness	Depressed mood	X	X	X
	Apparent sadness				
	Inability to feel				
	Suicide	Suicide	X		
Anxiety	Inner tension	Psychic anxiety	X	X	X
		Somatic anxiety			
Sexual function	NA	Loss of libido			
Appetite	Reduced appetite	Loss of appetite			
		Weight loss			
Sleep	Reduced sleep	Early insomnia			
		Middle insomnia		X	
		Late insomnia		X	
Functional status	Lassitude	Work and interests	X	X	X
		Agitation			
Ability to think	Concentration difficulties	Retardation	X		X
Physical symptoms	NA	Somatic symptoms general	X	X	X
Hypochondriasis	NA	Hypochondriasis			
General psychiatric distress	Pessimistic thoughts	Feelings of guilt	X	X	X
		Loss of insight			



(a) 9 items of the HAMD



(b) 8 items of the HAMD

Figure 3. Time course and score distribution for items from the HAMD. Upper panels show patterns in responders. Lower panels display patterns in non-responders. From left to right the following items are shown.

In (a): AGIT = *agitation*, ANXP = *anxiety psychic*, ANXS = *anxiety somatic*, APLS = *loss of appetite*, DPRM = *depressed mood*, FEGT = *feelings of guilt*, HYCD = *hypochondriasis*, INLS = *loss of insight*, INSE = *insomnia early*.

In (b): INSL = *insomnia late*, INSM = *insomnia middle*, LILS = *loss of libido*, RTRD = *retardation*, SOMG = *somatic symptoms general*, SUIC = *suicidal thoughts*, WINT = *work and interests*, WTLS = *loss of weight*

Endpoint selection and statistical power

A comparative analysis of the sensitivity of the MADRS, HAMD and the HAMD-subscales to detect treatment effect was performed using the mixed model for repeated measures. Based on the HAMD, the *p*-value for the drug-placebo difference following the 20 mg paroxetine dose in study 1 was 0.0573. In contrast, the *p*-value is lowered to 0.0232 when MADRS is used as endpoint. Such an increase in statistical sensitivity is further enhanced if the subscales of the HAMD are used (*data not shown*). In fact, subscale 1 (chapter 3) provides evidence that even the 30 and 40 mg treatment arms are significantly different from placebo, which none of the other endpoints were able to detect. For study 2, the significance levels were below 0.001 independent of the endpoint used.

To investigate the consequences of differences in sensitivity of the endpoints and its implication for the statistical power of the study, bootstrap simulations were performed using the data of studies 1 and 2. In figure 4, an example is shown of a treatment which was not superior to placebo (paroxetine 10 mg from study 1) and a treatment which was statistically superior to placebo (paroxetine arm from study 2). It is apparent that when no treatment effect is present, all endpoints perform equally, giving an indication of the control of type I error. As expected, enrolment of more patients does not increase the probability of a significant result. When a treatment effect is present, the MADRS is more powerful than the HAM-D$_{17}$ in detecting it. However, the one-dimensional subscales of

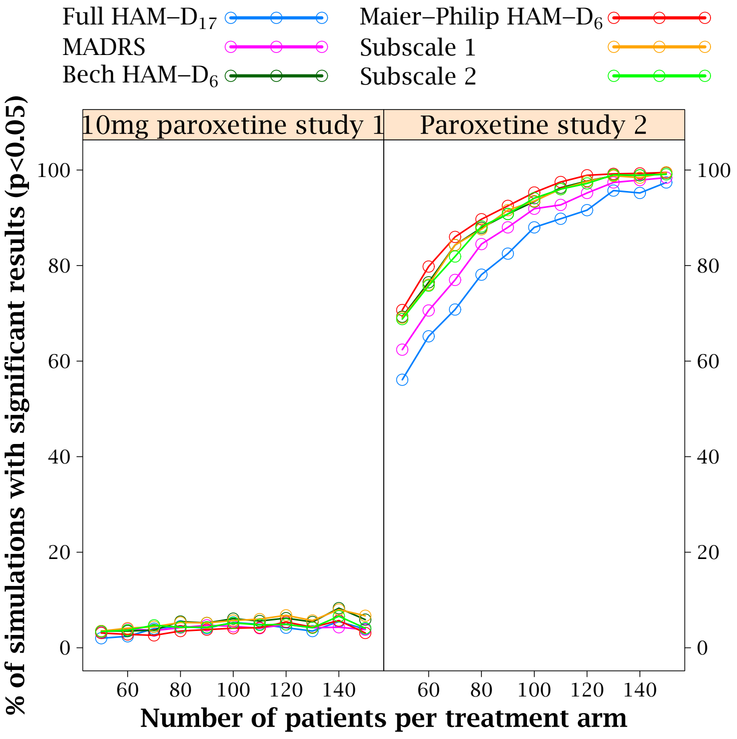


Figure 4. Results of the bootstrap simulations, showing the percentage of positive simulations out of 1000 *versus* the number of patients for the 10 mg paroxetine arm in study 1 and the paroxetine arm in study 2

the HAM-D_{17} outperform the MADRS.

The relationship between the number of enrolled patients and the power to detect treatment effect for all treatment arms in study 1 is shown in figure 5. The treatment effects range from none (10 mg paroxetine), via moderate treatment effects of the 30 and 40 mg paroxetine treatment arms to clearly significant treatment effects for the 20 mg treatment arm. The emerging trend is similar across all treatments (except for the 10 mg arm), with the HAM-D_7 response-based subscale (chapter 3) outperforming all other endpoints. The MADRS is the second-best endpoint, whereas the HAMD performs continuously worse.

For imipramine, the response-based HAM-D_7 (chapter 3) and the Bech HAM-D_6 were the most sensitive endpoints in study 2, followed by the other HAM-D_7 subscale proposed in chapter 3 and the MADRS (*data not shown*).

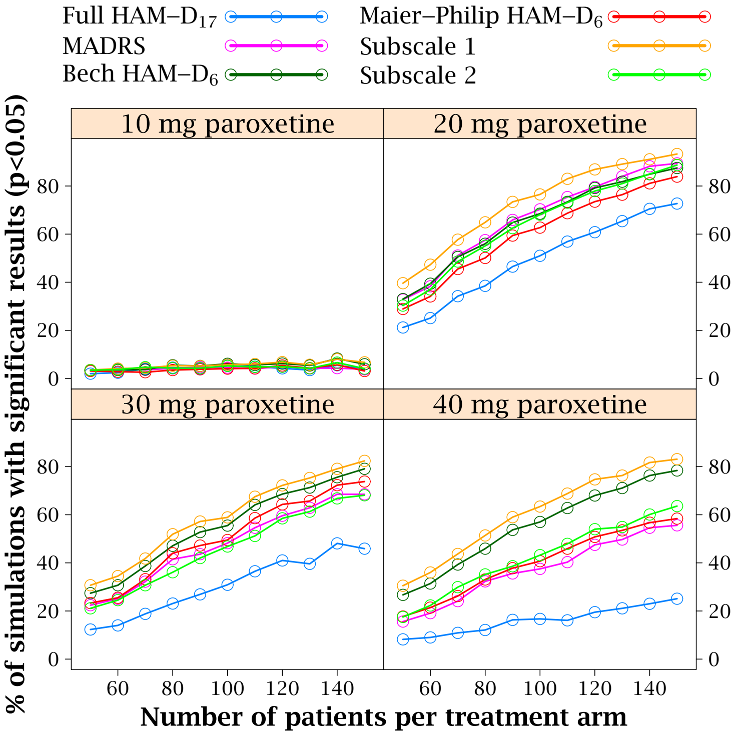


Figure 5. Results of the bootstrap simulations, showing the percentage of positive simulations out of 1,000 *versus* the number of patients in the treatment groups for study 1

DISCUSSION AND CONCLUSION

Sensitivity of clinical endpoints

Since the MADRS was designed to monitor change upon treatment, it is reassuring to find that nearly all individual items of the MADRS are sensitive to response, irrespective of treatment allocation. This contrasts to the results of a similar investigation into the items of the HAMD (chapter 3). Within these items large differences were found in their sensitivity to drug response. Indeed, some items, such as *loss of insight* and *loss of weight*, did not show any difference between the responder and non-responder populations. A comparison between the items of the MADRS and those items of the HAMD scale with corresponding content revealed that no major differences could be discerned. This is of particular interest, since it suggests that the MADRS and the selected items of the HAMD represent the same dimension of major depressive disorder. Furthermore, most of the items of the HAMD scale with corresponding value in the MADRS have been included in the subscales which were suggested previously. The domain of physical symptoms was represented in all HAMD subscales by *somatic symptoms general*, but it is not included in the MADRS (table 1). Therefore, a comparison between the MADRS, HAMD and the subscales of the HAMD with respect to sensitivity to treatment effect is of great interest.

Several authors have investigated the sensitivity to treatment effect of the MADRS and HAMD. Khan *et al.* (2004) investigated the correlation between the HAMD, MADRS and clinical global impression - severity (CGI-S) in a total of 347 patients. Their conclusion is that these endpoints, as was to be expected, are highly correlated. Also, comparing the effect sizes obtained using each endpoint in an LOCF analysis they conclude that there is no difference between these endpoints in sensitivity to detect drug effect. Although this study included patients treated on a single site, the reduced number of patients as compared to our analysis and the use of LOCF, rather than more appropriate statistical tools, may have biased the results. Moller (2001) has investigated the HAMD, MADRS and Bech-Rafaelsen melancholia scale (BRMS) using ’criteria of adequacy’, such as discriminating power, distribution of sum scores, internal consistency, content validity, homogeneity and transferability. Their analysis shows that MADRS and BRMS are superior to the HAMD, and they suggest comparing these endpoints with respect to sensitivity to drug effect. Carmody *et al.* (2006) have used item response theory models to compare the HAMD, its subscales and the MADRS. Similarly to the conclusions from the aforementioned publications, their results show superiority of the HAMD subscales and the MADRS and conclude that these scales may be better suitable for use as clinical endpoints in depression trials.

On the other hand, from a drug development point of view, the use of clinical scales remains a necessary evil. Compared to infectious diseases, where a clear difference exists between signs (bacterial colony growth) and symptoms (fever, erythema, pain), the distinction between signs and symptoms is much vaguer in psychiatry. Clinical scales perform a weighted assessment of a limited subset of the multitude of symptoms which constitute the syndrome of depression, and consequently remain far away from the underlying pathology. Based on the findings presented in table 1 and figure 2, the main difference between the HAMD and MADRS is that the latter only contains items which appear to be sensitive to response. Even so the MADRS does not encompass other potentially important domains of depression which are included in the HAMD (i.e., sexual functioning, physical symptoms and hypochondriasis). In fact, the item *somatic symptoms general* which captures the domain of physical symptoms was shown to be sensitive to response in our previous investigation and is also included in the Bech HAM-D$_6$.

It is evident that the manifestation of response also depends on the phenotype and on the severity of disease. It is conceivable that the items representing the domain of 'mood' will be relatively more important in melancholic patients than in other subgroups. In this sense, one must consider that the degree of severity may ultimately determine which domains are present (e.g., occurrence of physical symptoms). In many areas of medicine, clinical endpoints exist that are known to be related to specific mechanisms, usually a well-characterised receptor system. Examples of this link include the relation between the GABA-ergic receptor system and sedation (Tuk *et al.*, 1997) as well as the relation between the dopaminergic activity and extrapyramidal symptoms in Parkinson's disease (Bermejo *et al.*, 2007; Trosch, 2004). The endpoints used in depression are not *directly* linked to any of the mechanisms of action targeted by antidepressant drugs clinically available at present. This may indicate that the monoamine deficiency theory, upon which most of the existing antidepressants have been developed, may not be the ultimate explanation of the cause of mood disorders.

The availability of objective endpoints closely associated with pathways of disease would facilitate characterisation of drug effect and consequently improve the assessment of efficacy. Whilst such an endpoint requires advancement in the understanding of aetiology and disease progression, an opportunity exists to explore the distinct domains of disease. One could anticipate that specific receptor systems may be linked to specific domains. Another interesting concept is the development of composite scales, which combine subjective symptoms with other potential descriptors or biomarkers of disease severity, such as serotonin and cortisol levels or imaging results. One successful example of this approach is the recent introduction of the disease activity score (DAS) for the evaluation of efficacy in rheumatoid arthritis (van der Heijde *et al.*, 1993).

Implications of differences in sensitivity for statistical power & effect size

Since two studies were available in our database in which the HAMD and the MADRS were recorded concomitantly as endpoint, we have decided to investigate the sensitivity to change and treatment effect of these endpoints. The statistical model (MMRM) has been used in conjunction with a bootstrap procedure in a previous investigation (chapter 3). It takes into account all observations without the necessity to resort to a last observation carried forward (LOCF) approach, yielding unbiased estimates in the presence of data missing at random (Mallinckrodt *et al.*, 2004). Because the datasets can be created with any number of patients, it was also possible to investigate differences in sensitivity between endpoint in studies which have highly significant results across all endpoints. This is illustrated by study 2 in our investigation, since comparison of the *p*-values does not show a difference between endpoints (all $p<0.001$). In contrast, the bootstrap procedures clearly reveal that a difference exists between the MADRS, HAMD and its subscales.

It is important to note that we have chosen to use the same response criterion irrespective of the endpoint used, since a patient should be a responder or non-responder, irrespective of the selected endpoint. Given the acceptance of the HAMD as the gold standard, we have applied the HAMD response criterion also for the graphical analysis of the items of the MADRS. The results of our analysis show that the MADRS is more sensitive to treatment effect than the HAMD, although some of the HAMD subscales are more sensitive than the MADRS. This is especially true for the response-based subscale (HAM-D_7) proposed previously by our group (chapter 3), which was consistently one of the most sensitive subscales. Another important observation is that the subscales do not behave differently from the full HAMD on MADRS with respect to negative treatment effects, as illustrated by the 10 mg paroxetine treatment arm in study 1. This is critical, since an increased sensitivity to detect treatment effect should not be gained at the expense of an increased false positive rate.

Of course, the consequences of assessing treatment effect based on endpoints with different sensitivity depend on treatment effect size. Such differences however have major impact on study size, i.e., the number of patients required to achieve appropriate statistical power. This can be seen by drawing a vertical line in figure 2, for example at a power of 80%. If HAMD is used as endpoint, not even a treatment arm with 150 patients suffices to warrant equivalent power. In contrast, when MADRS is used as endpoint, 110-120 patients are enough. If one of the HAMD subscales is used, 100 patients are sufficient to reach 80% power to detect treatment effect. Consequently, the use of the response-based subscale 1 or MADRS would have led to different conclusions about the effect of the 40 mg treatment arm in study 1.

The implications of such differences in the sensitivity of the clinical scales are far reaching. Given the expected small differences between treatment arms in non-inferiority trials and head-to-head comparisons (i.e., small effect size), it becomes obvious how detrimental the selection of HAMD as endpoint really is. Analysis of such trials will certainly lead to inaccurate conclusions about treatment effect. Moreover, the selection of a more sensitive endpoint may unravel dose-response relationships. This is a critical issue in dose ranging pivotal trials (phase IIb), when dose-response curves ought to be established. The use of an endpoint with low sensitivity will obscure any such relationship.

In conclusion, the MADRS contains only items that are sensitive to response. Use of the MADRS as endpoint will increase the sensitivity to detect treatment effect and therefore allow enrolment of fewer patients. Yet, subscales of the HAMD seem to be even more sensitive to treatment effect than the MADRS. The selection of these subscales as primary endpoints in clinical trials could save over a third in patients compared to full the HAMD whilst keeping the same level of statistical power.

REFERENCES

Bech P and Rafaelsen OJ (1980) The use of rating-scales exemplified by a comparison of the Hamilton and the Bech-Rafaelsen melancholia scale. *Acta Psychiatr Scand* **62**:128–132.

Bermejo PE, Ruiz-Huete C, and Terron C (2007) Relationship between essential tremor, Parkinson's disease and dementia with Lewy bodies. *Rev Neurol* **45**:689–694.

Carmody TJ, Rush AJ, Bernstein I, Warden D, Brannan S, Burnham D, Woo A, and Trivedi MH (2006) The Montgomery Asberg and the Hamilton ratings of depression: A comparison of measures. *Eur Neuropsychopharmacol* **16**:601–611.

Dunner DL and Dunbar GC (1992) Optimal dose regimen for paroxetine. *J Clin Psychiatry* **53**:21–26.

Faries D, Herrera J, Rayamajhi J, DeBrota D, Demitrack M, and Potter WZ (2000) The responsiveness of the Hamilton depression rating scale. *J Psychiatr Res* **34**:3–10.

Feighner J, Cohn J, Fabre LF J, Fieve R, Mendels J, Shrivastava R, and Dunbar G (1993) A study comparing paroxetine placebo and imipramine in depressed patients. *J Affect Disord* **28**:71–79.

Fitzgerald J (2007) Description of the Hamilton depression rating scale (HAMD) and the Montgomery-Asberg depression rating scale (MADRS). *http://www.fda.gov/ohrms/dockets/AC/07/briefing/2007-4273b1_04-DescriptionofMADRS-HAMDDepressionR(1).pdf*.

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

Hamilton M (1967) Development of a rating scale for primary depressive illness. *Brit J Soc Clin Psychol* **6**:278–296.

van der Heijde DMFM, van 't Hof M, van Riel PLCM, and van de Putte LBA (1993) Development of a disease-activity score based on judgment in clinical practice by rheumatologists. *J Rheumatol* **20**:579–581.

Jonsson E (2004) Graphical display of population data. *PAGE* **13**.

Khan A, Brodhead AE, and Kolts RL (2004) Relative sensitivity of the Montgomery-Asberg depression rating scale, the Hamilton depression rating scale and the clinical global impressions rating scale in antidepressant clinical trials: a replication analysis. *Int Clin Psychopharmacol* **19**:157–160.

Khan A, Leventhal RM, Khan SR, and Brown WA (2002) Severity of depression and response to antidepressants and placebo: An analysis of the Food and Drug Administration database. *J Clin Psychopharmacol* **22**:40–45.

Maier W and Philipp M (1985) Comparative-analysis of observer depression scales. *Acta Psychiatr Scand* **72**:239–245.

Mallinckrodt C, Clark W, and David S (2001) Accounting for dropout bias using mixed-effects models. *J Biopharm Stat* **11**:9–21.

Mallinckrodt C, Kaiser C, Watkin J, Molenberghs G, and Carroll R (2004) The effect of correlation structure on treatment contrasts estimated from incomplete clinical trial data with likelihood-based repeated measures compared with last observation carried forward ANOVA. *Clin Trials* **1**:477–489.

Moller H (2001) Methodological aspects in the assessment of severity of depression by the Hamilton depression scale. *Eur Arch Psychiatry Clin Neurosci* **251 Suppl 2**:II13–II20.

Montgomery SA and Asberg M (1979) A new depression scale designed to be sensitive to change. *Br J Psychiatry* **134**:382–389.

O'Sullivan RL, Fava M, Agustin C, Baer L, and Rosenbaum JF (1997) Sensitivity of the six-item Hamilton depression rating scale. *Acta Psychiatr Scand* **95**:379–384.

R Development Core Team (2007) *R: A Language and Environment for Statistical Computing*, R

Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Santen G, Gomeni R, Danhof M, and Pasqua OD (2008) Sensitivity of the individual items of the hamilton depression rating scale to response and its consequences for the assessment of efficacy. *J Psychiatr Res* **doi:10.1016/j.jpsychires.2007.11.004**.

Trosch RM (2004) Neuroleptic-induced movement disorders: Deconstructing extrapyramidal symptoms. *J Am Geriatr Soc* **52**:S266–S271.

Tuk B, Oberye JJL, Pieters MSM, Schoemaker RC, Kemp B, vanGerven J, Danhof M, Kamphuisen HAC, Cohen AF, Breimer DD, and Peck CC (1997) Pharmacodynamics of temazepam in primary insomnia: Assessment of the value of quantitative electroencephalography and saccadic eye movements in predicting improvement of sleep. *Clin Pharmacol Ther* **62**:444–452.