Chapter 3

Sensitivity of the individual items of the Hamilton depression rating scale to response and its consequences for the assessment of efficacy

Gijs Santen, Roberto Gomeni, Meindert Danhof, Oscar Della Pasqua

*Journal of Psychiatric Research, In Press*
ABSTRACT

The Hamilton depression rating scale (HAM-D_{17}) has been the gold standard in depression trials since its introduction in 1960 by Max Hamilton. However, several authors have shown that the HAM-D_{17} is multi-dimensional and that subscales of the HAM-D_{17} outperform the total scale.

In the current study we assess the sensitivity of the individual HAM-D_{17} items in differentiating responders from non-responders over the typical treatment period used in clinical efficacy trials. Based on data from randomised, placebo controlled trials with paroxetine, a graphical analysis and a statistical analysis were performed to identify the items that are most sensitive to the rate and extent of response irrespective of treatment. From these analyses, two subscales consisting of seven items each were derived and compared to the Bech and Maier & Philip subscales using a linear mixed-effects modelling approach for repeated measures. The evaluation of two clinical trials revealed sensitivity comparable to the existing subscales. Using a bootstrap technique we show that the subscales consistently yield higher statistical power compared to the HAM-D_{17}, although no subscale consistently outperforms the others.

In conclusion, this study provides further evidence that not all items of the HAM-D_{17} scale are equally sensitive to detect responding patients in a clinical trial. A HAM-D_{7} subscale with higher sensitivity to drug effect is proposed consisting of the HAM-D_{6} and the suicide item. This response-based subscale increases the signal-to-noise ratio and could reduce failure rate in efficacy trials with anti-depressant drugs.

INTRODUCTION

An important problem in clinical trials with antidepressant drugs is the high failure rate in the assessment of clinical efficacy. Even in studies in which marketed antidepressants are administered at efficacious doses, failure rates of up to 50% are observed (Khan et al., 2002). Such a high failure rate may be due to several factors, among which 1) an inadequately powered study design, 2) the disease process itself, which is characterised by substantial variability, or 3) the sensitivity of the endpoint used in the studies. Thus far, limited quantitative research has been performed on the sensitivity and specificity of the clinical endpoint to the pharmacological effect over the treatment period. The current investigation was conducted to evaluate the influence of the latter on the estimation of treatment effect.

The Hamilton depression rating scale (HAM-D) has been the gold standard in depression trials after its introduction in 1960 by Max Hamilton (Hamilton, 1960). Since then, numerous authors have investigated the dimensionality of the scale and demonstrated that it is multidimensional (Bech et al., 1981; Moller, 2001). Others have evaluated its sensitivity to drug effect relative to other scales, such as the Montgomery-Asberg depression
rating scale (MADRS) (Montgomery and Asberg, 1979; Khan et al., 2004) and the Bech-Rafaelsen melancholia scale (MES) (Bech and Rafaelsen, 1980). Faries (Faries et al., 2000) has shown that a number of published one-dimensional subscales outperform the total HAM-D17 in sensitivity to drug effect and that the effect size of all published placebo-controlled trials with fluoxetine increases upon the use of these subscales as primary endpoint. In fact, the change in effect size was shown to be large enough to consider assigning one third less patients to studies whilst maintaining the pre-specified level of statistical power. Furthermore, Bagby and colleagues have written a review on the use of the HAMD (Bagby et al., 2004), which has been discussed extensively in a series of letters to the author (Bech et al., 2005; Carroll, 2005; Corruble and Hardy, 2005; Hsieh and Hsieh, 2005; Licht et al., 2005). The important conclusion from this paper is that the unwanted characteristics of the HAM-D17 warrant the development of a new gold standard.

In a recent review Bech (2006) discusses these issues and reaches the conclusion that the use of subscales as endpoint in an evaluation eliminates the confounding influence of non-specific items in the HAMD.

Originally, Hamilton (1960) did not intend the HAM-D17 to be used to monitor changes due to treatment effect. Rather its use was meant to characterise a depression state. It is thus possible to define clinical response during the course of a clinical trial based on the HAM-D17 under the assumption that a patient has reached some kind of steady-state. A frequently used definition for response is a decrease of 50% from baseline in total HAM-D17. Considering that the disease state information in the HAM-D17 is unbiased, the objectives of this investigation are to determine the sensitivity of the individual items of the HAM-D17 to clinical response over time and to develop a new subscale including only items that show a distinct pattern between patients who respond and patients who do not respond, irrespective of the treatment received during the trial. The impact of subscales on group size and statistical power will be assessed by a linear mixed-effects modelling approach for repeated measures (MMRM) on observed cases (OC) (Mallinckrodt et al., 2004). This method allows handling of missing data without the necessity to use the last observation carried forward (LOCF) approach. In contrast to LOCF, MMRM warrants unbiased results in the presence of data missing at random (MAR). Bootstrap methodology will then be used to explore the consequences of a reduction in the number of the patients in so-called Proof-of-Concept studies during early clinical development.

METHODS

Study data

Data from two clinical 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 frequency of clinical visits, total duration, well-defined criteria regarding patient population, design and dosing regimen. Patients should be diagnosed with MDD 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. Study 1 (phase II) was performed according to a double-blind, randomised, placebo-controlled design in which four fixed doses of paroxetine were investigated (Dunner and Dunbar, 1992). In this study, 50 patients were enrolled in the placebo arm and 100 patients in each active treatment arm. HAM-D assessments were carried out at baseline and at weeks 1, 2, 3, 4, 6, 9 and 12 after start of treatment. Study 2 (phase III) was also performed according to a double-blind, randomised, placebo-controlled design in which the efficacy of two different formulations of paroxetine was evaluated in an escalating dose design (Golden et al., 2002). A total of 315 patients were evenly enrolled across three arms. The HAM-D was assessed at baseline and at weeks 1, 2, 3, 4, 6, 8 and 12 after start of treatment. In one study, the HAM-D_{21} was used as endpoint. We have elected to use only the first 17 items in our analysis so that emerging subscales could be used in studies measuring the HAM-D_{17}, as defined by the revised rating scale HAM-D_{17} (Hamilton, 1967). Further details on the patient population and the study design are available in the original publication of the study results.

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. The only psychotropic co-medication allowed during treatment was chloral hydrate. In study 1, only up to four consecutive doses could be used during the first 2 weeks of the study. In study 2 such a restriction was not found in the protocol or report, but only 7.4% of patients made use of chloral hydrate. With regard to dropout, there were no significant differences in the HAM-D_{17} values of patients who dropped out, nor were the dropout times different between active or placebo treatment arms. An overview of the fraction of patients (%) remaining in the trial for each week and treatment arm is shown in figure 1.

**Subscale identification**

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 the HAM-D_{17} at baseline at any time during the trial. This definition of response is historically used, see for example (Tedlow et al., 1998). Dichotomisation and subsequent pooling of the data was performed after a preliminary evaluation showed no differences in the time course of response between treatments or any disparity in the time course of the individual HAM-D_{17} items across treatment groups in responders and non-responders. In principle, as per study protocol, each patient had a total of eight observations. All observations were grouped by week of visit and the time course of response was then analysed by showing the proportion of patients scored with each possible value for the individual item (Jonsson, 2004). Observations in week 8 in study 2 were grouped with the observations in week 9 in study 1. This procedure enabled extraction of information on three different levels or dimensions. First, it provides evidence of the time-dependence for the onset and maintenance
Sensitivity of the individual items of the Hamilton depression rating scale to response and its consequences for the assessment of efficacy

49

<table>
<thead>
<tr>
<th>Item</th>
<th>Study 1</th>
<th>Study 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Placebo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paroxetine 10 mg</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paroxetine 20 mg</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paroxetine 30 mg</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paroxetine 40 mg</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paroxetine CR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paroxetine IR</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1. Fraction of patients (%) remaining in the trial for each week and treatment arm in study 1 (left panel) and study 2 (right panel)

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. The influence of baseline HAM-D17 on the patterns of response was also evaluated by plotting the time course of each individual item separately for patients with low (≤ 22) and high baseline values (> 22).

Following the aforementioned data clustering, two approaches were used to evaluate individual items and potentially identify subscales that can better describe treatment effect, namely, a graphical analysis based on pattern detection throughout the course of treatment and a post-hoc statistical analysis based on the degree of response at the end of treatment. These were aimed at exploring the rate and extent of response during treatment, respectively.

All graphical analyses were performed in the language and environment for statistical computing R (R Development Core Team, 2007). HAM-D17 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 a time course that is clearly different between responders and non-responders. In contrast, slightly sensitive and insensitive items were those showing modest or minor differences, respectively. The items considered sensitive to response were grouped into a subscale...
To investigate whether the degree of sensitivity in individual items also varies with respect to the extent of response at the end of treatment, a Fisher exact test was used to assess differences in the proportion of responder and non-responders showing the lowest scores for each item at week 12. These items were then ranked in order of significance and a subscale was constructed based on these results (subscale 2).

Consequences of the use of subscales as a measure of response

Differences in the sensitivity of endpoints have direct implications for study power and consequently for the required study population size. The consequences of the use of subscales 1 and 2 as a measure of treatment response were assessed in the context of statistical data analysis in clinical trials, as compared to the HAM-D_{17}. Although it has been suggested that the analysis of the HAM-D_{17} and its subscales as a continuous variable should be abandoned in favour of an approach which focuses on the percentage patients in different disease categories (Bech et al., 1984), we have chosen a data analysis method that is currently considered state-of-the-art in depression trials. Therefore, 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 total HAM-D_{17}, the derived subscales and the published Bech et al. (1981) and Maier and Philipp (1985) 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 changes from baseline data. Baseline HAM-D_{17}, 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 $\alpha=0.05$ was used to establish the significance of treatment effect, which was determined as an average over all weeks.

In addition, a bootstrap analysis was performed to determine the power conditional on the data observed for each endpoint of interest. Since this method is less sensitive to the composition of the original datasets, accurate conclusions can be made about the relative sensitivity of the subscales as compared to the HAM-D_{17}. The bootstrap analysis consisted of re-sampling 1000 new populations with a size between 50 and 150 patients from the original studies. The replicated data sets were subsequently analysed using the MMRM approach in SAS. Although uneven randomisation occurred in study 1, equal group sizes were simulated for the purpose of bootstrapping.
RESULTS

The response patterns for each of the 17 items are presented in figures 2, 3 and 4, according to their classification into sensitive, slightly sensitive and insensitive, respectively. The graphical analysis reveals specific patterns and differences between individual items in distinguishing responders from non-responders. Some items, such as depressed mood, were very sensitive to response as indicated by clearly distinct time courses between responders and non-responders (figure 2). Other items, such as loss of weight were insensitive to response (figure 3). Interestingly, insensitive items seem to lack a time-dependent pattern, which is present in most sensitive items and to a lesser extent in slightly sensitive items (figure 4).

To investigate a possible effect of baseline HAM-D17 on the response patterns, separate graphs were created for patients with low and high baseline values (HAM-D17 ≤ 22 versus HAM-D17 > 22). No differences were found between these two groups (data not shown).

From the graphical evaluation, the seven items identified as sensitive to response over time (depressed mood, feelings of guilt, suicide, work and interests, retardation, anxiety psychic and somatic symptoms general) were grouped into a new response-based subscale (subscale 1).

In parallel, evaluation of the extent of response at the end of treatment showed that

![Figure 2](image_url)

**Figure 2.** Time course and score distribution for items from the HAM-D17 that were identified as sensitive to response. Upper panels show patterns in the responder subpopulation. Lower panels display patterns in the non-responder subpopulation. Abbreviations: ANXP=anxiety psychic, DPRM=depressed mood, FEGT=feelings of guilt, RTRD=retardation, SOMG=somatic symptoms general, SUIC=suicidal thoughts, WINT=work and interests.
Figure 3. Time course and score distribution for items from the HAM-D$_{17}$ that were identified as insensitive to response. Upper panels show patterns in the responder subpopulation. Lower panels display patterns in the non-responder subpopulation. Abbreviations: AGIT=agitation, HYCD=hypochondriasis, INLS=loss of insight, WTLS=loss of weight.

five out of the seven items selected graphically were among those identified with the highest level of statistical significance. The remaining two items from subscale 1, suicide and retardation, showed slightly lower statistical significance. Such an overlap provided an initial indication that sensitive items can, in fact, capture information about the rate and the extent of response. The seven most sensitive items were then grouped into another subscale (subscale 2). The total number of items selected was such to allow direct comparison of the performance of both scales with similar dimensionality. The only divergence between the two subscales was the identification of late insomnia (insomnia late) and maintenance insomnia (insomnia middle) instead of suicidal thoughts and retardation as sensitive items.

Evaluation of the sensitivity of the endpoints to drug effect was performed by re-analysing the original clinical data and comparing both subscales with the total HAM-D$_{17}$ using the MMRM. For study 1, the contrast for 20 mg paroxetine against placebo turned out not to be statistically significant ($p=0.0566$) on the change from baseline using the full HAM-D$_{17}$. Using the two subscales, the same analysis resulted in statistically significant treatment effects ($p=0.0154$ and $p=0.025$, respectively). The contrast for 10 mg paroxetine remained statistically not significant irrespective of the endpoint used, whereas the $p$-value for 30 mg paroxetine changed from 0.1799 to 0.043 and 0.0837 and for 40 mg paroxetine from 0.4238 to 0.0499 and 0.1086 for the HAM-D$_{17}$, subscale 1 and subscale 2,
Sensitivity of the individual items of the Hamilton depression rating scale to response and its consequences for the assessment of efficacy

Figure 4. Time course and score distribution for items from the HAM-D17 that were identified as insensitive to response. Upper panels show patterns in the responder subpopulation. Lower panels display patterns in the non-responder subpopulation. Abbreviations: ANXS = anxiety somatic, APLS = loss of appetite, INSE = insomnia early, INSL = insomnia late, INSM = insomnia middle, LILS = loss of libido respectively. The use of the Bech and Maier subscales for 20 mg paroxetine resulted in a decrease in statistical significance to 0.0285 and 0.0379. These subscales did not perform as well for the higher doses of paroxetine.

In study 2, in which two different paroxetine formulations were evaluated, the contrast of formulation 1 (modified release) to placebo remained similar and highly significant irrespective of the subscale used. The contrast of formulation 2 (immediate release) to placebo changed from 0.0595 for the HAM-D17 scale to 0.0063 and 0.0199 using subscales 1 and 2, respectively. Similarly, the Bech and Maier subscales reduced the p-values of formulation 2 to 0.0092 and 0.0016. Based on the aforementioned findings, an attempt was made to evaluate the consistency of the differences in the sensitivity of the subscales. The analysis was therefore repeated for various clinical studies in which paroxetine and other investigational compounds were administered, confirming the increase in sensitivity to treatment effect observed in study 1 and 2 (data not shown).
Figure 5. Results of the bootstrap analysis, showing the percentage of simulations in which the treatment effect reaches statistically significant differences from placebo (p<0.05) against the number of patients per treatment group.

Since lower probability values for statistical significance are obtained using the proposed subscales, it may well be possible to design trials with fewer patients while maintaining the same statistical power. We used therefore a bootstrap technique to generate new study populations and assess the power conditional on the observed data and sample sizes ranging from 50 to 150 patients. The results of these simulations are summarised in figure 5. For the sake of clarity, only one treatment-placebo difference per study is shown.

The differences for other treatment arms versus placebo were similar, except for the 10 mg paroxetine arm in study 1, for which statistically significant separation from placebo could not be achieved by any of the scales under evaluation.

DISCUSSION AND CONCLUSION

In the current investigation we have used an empirical approach to dissect the multidimensionality of HAM-D17 scale, showing that the ability of individual HAM-D17 items to discriminate between response and non-response is time-dependent and highly variable. Despite the slight variation in the ranking of the most sensitive items when considering rate (on a graphical basis) versus extent of response (using the Fisher exact test), these findings show that the statistical significance of a treatment effect is highly dependent on the sensitivity of the selected endpoint and that any meaningful conclusion about effect
Sensitivity of the individual items of the Hamilton depression rating scale to response and its consequences for the assessment of efficacy

size must take time into account, since the differences in response between drug and placebo change over time. In addition, we have not identified any differences in response pattern between patients with low and high baseline values. Our results seem to contrast with the data reported by Tedlow et al. (1998), who have implied that low baseline values are positively correlated with the probability of being a responder.

The current findings are pertinent to two major issues in clinical drug development. First, it may partly explain the high failure rate in proving treatment effect even when drugs with known efficacy are administered in a trial. Second, it may clarify some of the difficulties in demonstrating efficacy of new compounds and establishing the appropriate dosing regimen for specific mood disorders related to depression, for which HAM-D17 remains the gold standard. Moreover, such a lack of discriminatory power for response has further repercussions on the characterisation of placebo response and concentration-effect relationships. Recently, we have shown that the use of pharmacokinetic-pharmacodynamic (PKPD) modelling to characterise concentration-effect relationships and consequently optimise dose selection in clinical efficacy trials in early clinical development is only meaningful when appropriate measures or markers of disease are available (Franciosi et al., 2006; Maas et al., 2006). In these publications, the authors challenge the general assumption that diagnostic endpoints often associated with disease stationarity can be used to accurately assess drug effect under non-stationary conditions, as is the case during the course of treatment. Furthermore, the investigation by Maas et al. (2006) shows that symptom-based endpoints specific for diagnostic purposes can only be used as primary endpoint in clinical trials under the assumption of equal sensitivity to the onset and offset of disease. This assumption implies reversibility of the underlying disease processes as well as reversibility of the pharmacological action. Such a state of reversibility, however, is often not true, nor is the mechanism by which drugs exert their action in major depression and other depressive mood disorders.

Our approach differs from that of Evans et al. (2004) in that it does not use item response theory to determine the most sensitive items. A necessary assumption in item response theory is that the full scale properly captures the variable of interest. As discussed extensively in this paper, it is unlikely that the HAM-D17 is a good continuous measure for the severity of depression. Our approach only assumes that the HAM-D17 is a proper marker for the disease state, which is exactly its purpose. It is interesting to note that although the underlying assumptions are different, the items deemed most useful were similar. The only important difference are two items included in the subscales presented in this paper, which do not appear in the work of Evans et al. (2004), namely suicide and retardation.

It is also worth mentioning that 6 out of the 7 most sensitive items identified from the longitudinal response pattern analysis are also included in the HAM-D6 core subscale for depression (Bech et al., 1981), and 5 of them are included in the 6-item Maier and Philipp (1985), even though different selection criteria were used in these investigations. Intriguingly, the items that do not appear in any other published subscale are suicide in
Chapter 3

subscales 1 and *insomnia late & insomnia middle* in subscale 2. The first is of particular interest since depressed patients often suffer from suicidal ideation. Hence, a responder to treatment should endure less of such thoughts (Nierenberg *et al.*, 2001) and show lower scores for this item. It is remarkable to establish the sensitivity of this item, despite the fact that patients with a high likelihood of suicide are usually excluded from clinical trials. This exclusion criterion was also applied to the studies used in our investigation. Regardless of the low incidence of moderate and high scores, the time course of the *suicide* item is still markedly different between responders and non-responders. In fact, this is the first time this contrast is made; to date previous research has focused on discriminating between placebo and active treatment, in which case no differences are observed (Khan *et al.*, 2000). The impact of the addition of *suicide* on the scale’s sensitivity may be assessed by comparing subscale 1 and the Bech HAM-D6. Subscale 1 was shown to perform slightly better in the two studies reported in this paper and in additional efficacy trials with SSRIs and tricyclic antidepressants (*data not shown*). However, some care needs to be taken when applying subscale 1 to clinical trials in which the exclusion criteria for patients with suicidal ideation differ from the studies reported here. According to published reports linking paroxetine and the other SSRIs to suicidal *behaviour* (Healy, 2006; Lenzer, 2006), one could anticipate that any subscale including the *suicide* item might yield a smaller treatment effect for paroxetine, as compared to a subscale without this item. Our findings show however quite the opposite. As can be seen in figure 5, a significant difference exists between subscale 1 and the HAM-D6. Whilst we do not refute the possibility of a link between SSRIs and suicidal behaviour, this analysis shows that in responders there is considerable improvement on the *suicide* item during the course of therapy with paroxetine. One might argue therefore whether patients who develop suicidal behaviour are more likely to belong to the non-responder population, and hence have not benefited from treatment.

With respect to the items emerging from subscale 2, *insomnia late & insomnia middle*, it is known that early awakening, is a well known symptom in depression, often associated with anxiety (Taylor *et al.*, 2005). Since insomnia is a common finding in most patients with generalised anxiety disorder (GAD) (Culpepper, 2002), this item may be of value in the assessment of drugs for that specific indication. Furthermore, the overlap in item sensitivity identified at the end of treatment seems to provide evidence that the extent of response is correlated with the rate of response during the course of therapy. This may have implications for the use of subscales in interim data analysis.

Instead of combining all nine items arising from the proposed analyses, we have chosen to proceed with the evaluation of the MMRM approach using both HAM-D subscales, each containing 7 items. Either set revealed to be consistently more sensitive than the full HAM-D17 scale in detecting treatment effect if used as endpoint in the MMRM, indicating that these endpoints are a suitable measure of efficacy in anti-depressant trials. Interestingly, subscale 1, which is based on longitudinal response patterns, performed clearly better. Such a difference in performance is evident from both the bootstrap results, which
show that the use of a subscale could allow a reduction in group size while maintaining
the same statistical power and also from the \( p \)-values resulting from the study analysis
which are significantly lower for subscale 1 as compared to the full HAM-D\(_{17} \). The MMRM
approach and the simulation methodology used in this investigation were selected to ac-
count for missing data and reduce the bias from last observation carried forward (LOCF)
analysis, which is common practice in the statistical analysis of depression trials and of-
ten required by regulatory agencies. These techniques are therefore more accurate and
do better than the previous work by Faries et al. (2000). Most importantly, this method
allowed the estimation of the conditional power, which is a better measure to compare dif-
ferent endpoints accounting for the influence of the composition of the original datasets.

A comparison between the new subscales and the Bech and Maier subscales shows
that the \( p \)-values obtained for subscales 1 and 2 are lower or similar, suggesting greater
separation between placebo and active treatment effect. Elsewhere, the Bech HAM-D\(_{6} \) has
been shown to be superior to the full HAM-D\(_{17} \) for both typical and atypical depressed
patients (O’Sullivan et al., 1997). The bootstrap simulations show however that the Maier &
Philip subscale has a higher conditional power than the subscales suggested here when the
data from study 2 is used, although it performs considerably worse when the data from
study 1 is used. Yet, the purpose of this paper was not to devise a ‘better’ subscale than
those already available, but rather to show that the selection of items that are sensitive to
response (irrespective of treatment) results in higher estimates of treatment effect. These
observations highlight another issue around sensitivity and specificity of HAM-D\(_{17} \) scale,
namely whether the changes over time are disease-specific or can be associated to a drug’s
mechanisms of action. The evaluation of response patterns following administration of
different drug classes is currently being undertaken by our group and will be presented
in a separate publication.

It is apparent from the results of the MMRM analysis that the remaining items of
the HAM-D\(_{17} \) add considerable noise to the signal arising from response-sensitive items.
This also implies that summing up the scores does not provide a correct measure for
drug effect (Licht et al., 2005). Indeed, improvement over placebo became significantly
different when the new subscales were used. Based on the original data analysis, drug
effect in two treatment arms was not considered statistically different from placebo. Most
importantly, the increase in sensitivity achieved by the use of the subscales correlated
to dose level. As expected for this indication, the 10 mg paroxetine arm did not show
separation from placebo, whilst treatment response to the 30 and 40 mg paroxetine was
statistically different from placebo. In a previous report, a clear increasing dose-response
relationship was observed for the melancholic subpopulation enrolled into study 1 (Tignol
et al., 1992). This subset of patients consisted of about 50% of the total population. In
the full dataset we found that the 20 mg paroxetine arm was most different from placebo,
with 10 mg paroxetine not showing any difference and the 30 and 40 mg paroxetine arms
differing slightly less than the 20 mg paroxetine arm. It has to be noted that this trial
is relatively small, especially the placebo arm (n=50). Therefore these results are not
sufficient proof of the lack of a dose-response relationship. A similar inconsistency was reported for citalopram, for which comparable efficacy was observed for the 10 mg and 20 mg doses in the ITT population. In contrast, this discrepancy disappeared when the analysis was repeated in only severely depressed patients (Bech et al., 2004).

Although other reasons for study failure such as inadequate sample size, patient population and observation schedule remain as important as ever, our investigation shows that the use of change from baseline of the HAM-D$_{17}$ as primary endpoint does contribute to failures in the assessment of treatment efficacy. Another interesting point to be derived from this analysis is the time course of improvement observed within each individual item. Improvements are observed immediately after start of treatment in the response population as evidenced by the score patterns of the sensitive items. Previous investigations have also shown that there is no real evidence for the common assumption that a 3-5 week delay in improvement after start of treatment exists (Posternak and Zimmerman, 2005; Stassen and Angst, 1998). Unfortunately, in practice this delay in the onset of effect is still considered true and communicated to patients, medical students and health professionals.

In conclusion, this study adds further data to the body of evidence that the HAM-D$_{17}$ is not the most sensitive measure of treatment effect in depression trials. This is especially important in early clinical development studies, in which relatively small patient populations are included, increasing the risk of false negative results. The social and financial costs of false negative results are not acceptable in an indication which urges for medicines with a better efficacy profile. Furthermore, we show that a considerable reduction in population size can be achieved when the subscales are used due to an increase in statistical power to detect differences from placebo. Such a smaller group size does not only reduce the run time of a study, it is also limits the unnecessary exposure of patients to placebo treatment. From the two subscales we have devised, subscale 1, derived from graphical analysis of the longitudinal patterns of individual items, performs better in the detection of treatment effect than subscale 2. Therefore, we recommend the use of this response-based HAM-D$_{7}$ subscale as primary endpoint for the statistical analysis of efficacy data rather than the HAM-D$_{17}$.

REFERENCES


Sensitivity of the individual items of the Hamilton depression rating scale to response and its consequences for the assessment of efficacy


BMJ 332:1175.