Rating scales in depression: limitations and pitfalls

Per Bech, MD

Since the introduction of antidepressants to psychopharmacology in the 1960s, the Hamilton Depression Rating Scale (HAM-D) has been the most frequently used rating scale for depression. When used as a scale for prediction of outcome with antidepressants, the HAM-D, by its total score, has obtained limited use analogous to the Diagnostic and Statistical Manual of Mental Disorders. *4th ed* (DSM-IV) *diagnosis of major depression. Most research has been devoted to the use of the HAM-D to discriminate between placebo and active drugs or to show dose-response relationship in patients with major depression. An improvement in the total HAM-D score during a drug trial does not, however, in itself qualify the drug as an antidepressant, because the total score is not a sufficient statistic. The problem of statistical versus clinical significance when analyzing placebo-controlled trials, including dose-response relationship, is outlined, with the recommendation to use effect size statistics.*

 Dialogues Clin Neurosci. 2006;8:207-215.

Keywords: *depression rating scales; HAM-D; antidepressant; major depression; placebo-controlled trials*

Author affiliations: Psychiatric Research Unit, Frederiksborg General Hospital, Hillerød, Denmark

Address for correspondence: Prof Per Bech, Professor of Psychiatry, Psychiatric Research Unit, Frederiksborg General Hospital, 48, Dyrehavevej, Hillerød, Denmark (e-mail: pebe@fa.dk)

Depression rating scales were introduced into clinical psychiatry in the 1960s, with the advent of antidepressants such as imipramine and phenelzine.[1-3] In the early trials, both global improvement scales and the Hamilton Depression Rating Scale (HAM-D) were used. As discussed by Lam et al,[1] historically the use of depression symptom scales such as the HAM-D was not a routine aspect of patient care for frontline mental health clinicians. The present situation seems to be that we are facing two prototypes of clinicians, "Dr Gestalt," who uses a global clinical impression scale, and "Dr Scales, " who has incorporated the routine use of rating scales into daily clinical practice.[1]

When comparing Dr Gestalt with Dr Scales with respect to limitations and pitfalls in using depression rating scales, it seems appropriate to use the functional analysis proposed by Emmelkamp.[2] According to this proposal, we can refer to macroanalysis and microanalysis of rating scales. Macroanalysis focuses on the diagnosis of depression and thereby the prediction of treatment response, while microanalysis focuses on outcome measures of treatment. At the macroanalytic level, it is appropriate to discuss depression rating scales such as the HAM-D in comparison with a diagnostic system of mental disorders such as the *Diagnostic and Statistical Manual of Mental Disorders*, 4th ed *(DSM-IV)*,[3] while at the microanalytic level a direct comparison between Dr Gestalt and Dr Scales is relevant.

Macroanalysis

Emmelkamp[2] used the polythetic algorithms of the *DSM-IV* to illustrate the limitation of the clinical diagnosis of depression when developing treatment strategies for the patients. According to *DSM-IV*, in major depression five out of nine depression symptoms have to be

Selected abbreviations and acronyms

CGI	*Clinical Global Impression Scale*
DSM-IV	Diagnostic and Statistical Manual of Mental Disorders. *4th ed*
HAM-D	*Hamilton Depression Rating Scale*
MADRS	*Montgomery-Asberg Depression Rating Scale*
SSRI	*selective serotonin reuptake inhibitor*

present. This implies, as discussed by Emmelkamp, that totally different patients may fulfil these symptomatic requirements, because the fixed number of five items may refer to different items from patient to patient. Consequently, this heterogeneity has serious limitations for the predictive validity of the diagnosis concerning choice of treatment.

In 1979, the Montgomery-Asberg Depression Rating Scale (MADRS) was introduced into clinical psychiatry because the existing depression rating scales reflected " ... diagnostic features rather than being sensitive to change ... ."[4] Thus, the HAM-D was considered was considered by Montgomery and Asberg to be a diagnostic scale although Hamilton had designed it as a scale measuring the severity of depressive states and not the diagnosis.[5]

After 1980, with the introduction of the *Diagnostic and Statistical Manual of Mental Disorders*, 3rd ed *(DSM-III)*[6] the diagnosis of depression was symptom-based, but, as illustrated by Emmelkamp,[2] the algorithm of major depression is resistant to quantification.

Studies with the HAM-D have indicated that the HAM-D is not a unidimensional scale,[7] suggesting that the profile of factors, eg, suicidal behavior, anxiety-somatization, sleep, and appetite or weight loss should be used in a macroanalytic approach when developing a treatment strategy with antidepressants.

In the study by Montgomery and Asberg,[4] the item most sensitive to change during treatment was the sleep item; this may be explained by the antidepressants used in the analysis (amitriptyline, clomipramine, maprotiline, and mianserin). One of the limitations of depression rating scales as claimed by Montgomery and Asberg[4] was that they are only rarely consistent in finding differences between active drugs, even when the known mechanisms of action are different. However, in a judgment analysis it was found that clomipramine was superior to citalopram, but only on the item of sleep and not on the specific items of depression.[8] We can thus differentiate between sedative antidepressants such as amitriptyline, clomipramine, and mianserin (all antihistamines) and nonsedative antidepressants such as as citalopram or other selective serotonin reuptake inhibitors (SSRIs). In this context, the sleep and agitation factor on the HAM-D might become predictive of choice of antidepressants. However, Katz et al[9] have argued for also including factors such as somatization, hostility, and interpersonal sensitivity from the Symptom Checklist (SCL-90) when selecting the type of antidepressant. Likewise, the symptom of suicidal behavior should be analyzed separately when selecting the most appropriate treatment and care for the patient.

Macroanalyses of rating scales are rarely performed, but a multidimensional scale such as the HAM-D might give the clinician better information than the *DSM-IV* diagnosis of major depression when selecting the most appropriate antidepressant treatment for the individual patient. With the *DSM-IV* symptoms of depression it is possible to create a profile of a patient by the score on agitation versus retardation, suicidal behavior, sleep problems, and weight loss versus weight gain.

The only rating scales designed specifically to measure predictive validity of treatment by their total scores are the Newcastle Depression Scales (Newcastle 1965[10] and Newcastle 1971[11]). With the introduction of *DSM-III* and *DSM-IV*, the subdivision of depression into endogenous and reactive depression was deleted, and research on the Newcastle scales, which had been based on this concept, became very limited.

The various guidelines on how to use the different antidepressants with reference to treatment-specific algorithms are typically based on the safety of the drugs and the patient-specific history of treatment resistance, rather than on the *DSM-IV* diagnosis of major depression or on a score on a depression rating scale.[12]

Research on how to uncover medication history to help with the treatment decision has been very limited. Posternak and Zimmerman[13] have recently examined how accurately patients can recall prior treatments with antidepressants. The results showed that approximately 80% remembered monotherapy correctly, while only 25% recalled augmentation therapy correctly.

In the macroanalysis of the choice of treatment, it must therefore be concluded that rating scales with a factor profile such as the HAM-D seem to be superior to the *DSM-IV* diagnosis of major depression, but the *DSM-IV* depression symptoms individually can give important information about choice of treatment. However, when

making decisions about individual patient-specific treatments, the tolerability of the antidepressant plays an important role, as does the history of previous outcome, especially in regard to treatment resistance.

Microanalysis

According to Emmelkamp,[2] the microanalysis of a depression rating scale is mainly focused on the clinimetric analysis of outcome measurements of treatment. This type of analysis, as discussed by Faravelli[14] is based on certain assumptions which often involve pitfalls to such a degree that they can lead to "evidence-biased" rather than "evidence-based" psychiatry. The assumptions listed by Faravelli are:

- An illness is the sum of its symptoms;
- The symptoms are represented by the numbers associated with specific behaviors;
- Operations conducted statistically on these numbers reflect actual changes in the clinical reality;
- The relationship among numbers is represented by simple additive effect, regardless of reciprocal interaction.

These assumptions are the focus of the dialogue between Dr Gestalt and Dr Scales.[1] One of the aspects discussed by Lam et al[1] is that Dr Gestalt in his treatment may focus only on one symptom which might be misleading, while Dr Scales has a fuller picture of the patient's current state. From Faravelli's point of view, Dr Gestalt is a very experienced psychiatrist, while in Lam's discussion Dr Gestalt is as inexperienced as Dr Scales.

It is certainly a disadvantage to believe that the use of depression rating scales is an attempt to replace experienced psychiatrists by young and inexperienced clinicians in clinical trials. In this context it is important to be aware of the instructions for the Clinical Global Impression Scale (CGI) by Guy.[15] When using the CGI, the clinician has to make his or her assessment on the basis of previous experience with depressed patients. It is thus with reference to experience that the clinician should make the comparison with all the other severely depressed patients he or she has ever treated. In their daily routine, as stated by Hamilton,[16] experienced clinicians always perform a global rating when assessing a depressed patient's need for hospitalization or when deciding whether to discharge an inpatient.

The clinically most significant method for validating a depression symptom rating scale such as the HAM-D is to use experienced psychiatrists, both in the group of raters making the global assessment and in the group of raters making the rating scale assessment. This approach was analyzed by Bech et al[17] and showed that both groups of experienced psychiatrists were able to obtain an adequate interobserver reliability on the global assessment as well as in HAM-D ratings. An item analysis showed that only six of the 17 HAM-D items validly reflected the global assessment.[17] These six items (HAM-D_6) are shown in *Table I*. The three items listed at the top of *Table I* are the specific items of depression in accordance with *DSM-IV* and the *International Classification of Diseases*, 10th revision *(ICD-10)*.[18] This was supported by Hamilton in his last study,[19] in which he also demonstrated that the item of psychic anxiety is a specific item of depression. The remaining two items in *Table I* are "guilt feelings" and "psychomotor retardation." Guilt feelings are the specific item of negative thoughts which, according to Beck's cognitive model, are a central feature of depressive states.[20] Psychomotor retardation is the most specific observational symptom of depression, and in the Melancholia Scale (MES), which is a depression rating scale based on the HAM-D_6, the item "psychomotor retardation" has been subdivided into motor, verbal, intellectual, and emotional retardation.[21]

As discussed by Frances et al,[22] the items considered to be most specific for a disorder such as depression might have poor ability to discriminate this disorder from other

HAM-D items	HAM-D subscales		
	HAM-D_6 (20)	Maier subscale (31)	Core factor subscale (32)
Depressed mood	+	+	+
Work and interests	+	+	+
General somatic (tiredness)	+		
Psychic anxiety	+	+	
Guilt feelings	+	+	+
Psychomotor retardation	+	+	+
Psychomotor agitation		+	
Suicide			+
Clinimetric validity			
a) global impression by experienced psychiatrist	+	-	-
b) Psychometric analysis			
• Item response analysis	+	+	-
• Factor analysis	-	+	+

Table I. Specific depression subscales derived from the HAM-D by the microanalytic approach.

disorders, and the items that are most dicriminating may not be close to the core symptoms. The HAM-D_6 items in *Table I* are those that in the microanalytic sense are specific for antidepressant activity, while the items identified at the macroanalytic level to discriminate between treatments are, for instance, sleep, appetite, agitation, and suicidal behavior.

Table I shows the three most frequently used subscales for measuring antidepressant activity. The HAM-D_6 has been used in trials with fluoxetine,[23] citalopram,[24] escitalopram,[25] paroxetine,[26] and mirtazapine,[27] while the Maier subscale[28] and the core factor subscale[29] have recently been included in the duloxetine program.[30] The order of HAM-D items in *Table I* is listed according to their appearance in the depressive states when having taken into account the severity degrees of the individual items. To be additive in Farvelli's sense, the individual items of a rating scale must be consistently rank-ordered according to their relation to the severity of depressive illness. This implies that scoring of lower-prevalence items (low appearance) presupposes scorings on higher-prevalence items (high appearance). Thus, a score on guilt feelings or psychomotor retardation (which has low prevalence) has to be preceded by high scores on depressed mood and work and interests (which have the highest prevalence). The statistical analysis based on this criterion of additivity (ie, the total score being a sufficient statistic or unidimensionality of the scale items) is referred to as item response analysis.[26] The item of psychomotor agitation was excluded from the HAM-D_6 development by both the experienced psychiatrists[17] and by the item response theory model[26] because of a reciprocal interaction with the other items.

As indicated in *Table I*, the clinimetric background for the Maier subscale is an item response analysis which was performed in a study showing that the HAM-D_6, in contrast to the MADRS, was a unidimensional scale, and where the Maier subscale emerged as a by-product of the statistical analysis.[28] The core factor subscale was identified by an exploratory factor analysis by Cleary,[29] but has never been confirmed by other factor analyses. A recent comparison between HAM-D_6 and the Maier subscale[31] has shown that both scales were valid, while the CGI was unreliable. Although the theoretical score range of the HAM-D_6 goes from 0 to 22 and that of the Maier subscale from 0 to 24, the standardization of the two scales showed identical cutoff scores. Thus, a score above 10 on the Maier subscale indicates 18 on the HAM-D_{17} (moderate depression) and a score above 12 indicates 25 on the HAM-D_{17} (severe depression), while a score below 5 indicates 7 on the HAM-D_{17} (remission). As no patient can have a maximum score on both psychomotor retardation and psychomotor agitation, the Maier subscale should be considered having a practical score range corresponding to the HAM-D_6.

Neither in the MADRS nor in the Melancholia Scale (MES) is the item of psychomotor agitation included. Therefore, to develop a $MADRS_6$ subscale to cover the specific depression items according to *Table I*, only the HAM-D_6 is available.[24] The psychometrically most significant method for analyzing Faravelli's assumptions is the use of item response theory models.[26] By use of the nonparametric model of Mokken it has been shown that the $MADRS_6$ is also a unidimensional depression scale.[24] The $MADRS_6$ includes the corresponding HAM-D_6 items.

A major pitfall in a microanalysis of the HAM-D is the use of factor analysis to test Faravelli's assumptions. A comprehensive review by Bagby et al[7] has shown that factor analysis as used from 1980 to 2003 in many psychometric analyses of the HAM-D has identified quite different factor scores. As discussed elsewhere,[32] the clinimetric analysis of a rating scale should indicate to what extent the total score is a sufficient statistic by considering both the individual items of the scale and the population under examination.

When trying to define the antidepressant effect of a drug, Prien and Levine[33] concluded that a greater improvement in total HAM-D scores does not necessarily indicate antidepressant action ("... assume that a group treated with an experimental drug shows significantly more improvement than a group treated with placebo on the factors of anxiety, somatization or sleep disturbances and no significant change on other factors. These changes, by themselves, should not qualify the drug as an antidepressant..."[33]).

Another major pitfall to be considered is the use of several depression scales in the same trial without clearly indicating a priori which of them has been determined to be the the primary measure of antidepressant effect. To avoid this problem, a researcher should always use the specific items of depression, eg, the HAM-D_6 or the $MADRS_6$, as the primary efficacy measure. When determining clinically significant antidepressant effect, it is recommended to use standardized effect size statistics.[34] These statistics examine the reduction of rating scale scores from baseline to end point (mean scores) for both

active drug and placebo in relation to the pooled standard deviation of the two treatments. Thus, if the baseline score is 24 for both treatments, but the change score is 14 for the active drug while it is 10 for the placebo, and if the pooled standard deviation is 8, then the effect size is 4/8 or 0.50. In clinical trials with antidepressants an effect size of 0.40 or higher is considered a clinically significant response criterion.[35] This equals a 20% advantage of the active drug over placebo by using either a global impression score of very much and much response[36] or a 50% reduction in baseline rating scores on the HAM-D.[23]

Illustrating antidepressant effect, as shown in *Figure 1,* is yet another difficult area. Because both groups of patients, ie, on active drug treatment as well as on placebo treatment, exceed 100 subjects, a small statistically significant difference will be found. In the example illustrated in *Figure 2*, it is obvious that the effect of escitalopram is of clinical significance (effect size >0.40) in depressed patients after only 4 weeks.

In dose-response trials, the HAM-D_6 and the $MADRS_6$ were much more sensitive than the full versions of the respective scales, ie, HAM-D_{17} and $MADRS_{10}$.[23,37] Both the HAM-D_6 and the Maier subscale obtained an effect size of approximately 0.50 for venlafaxine and 0.40 for fluoxetine in placebo-controlled trials in patients with major depression, while the HAM-D_{17} even for venlafaxine, obtained an effect size of below 0.40.[38]

In a comparison of most of the placebo-controlled trials of SSRIs in patients with major depression[39] it was found that the HAM-D_{17} was used more frequently than the $MADRS_{10}$. As no difference was seen between the two scales in differentiating between active drug and placebo, only the HAM-D_{17} results were considered.[39]

The correct use of depression rating scales in clinical trials of antidepressants is, as illustrated in *Figure 2*, to indicate the effect size of the specific items of depression and to accept an effect size of 0.40 or higher as being the clinically significant effect. The current tradition of including at least two depression rating scales without focusing on the specific items of depression seems to constitute a "scientific wrapping" with which the companies decorate their antidepressants, eg, in a figure analogous to *Figure 1*. This industry habit of "dressing" antidepressant activity does now also include the use of the Hamilton Anxiety Scale (HAM-A) to show the antianxiety activity of an SSRI. The 14-item version of the HAM-A[40] includes an item of depressed mood. However, when using the HAM-A to indicate an effect on generalized anxiety, only its specific items should be used.[41] The HAM-A subscale with the six specific items of generalized anxiety is shown in *Table II*.[42]

When evaluating the antidepressant activity of new drugs in placebo-controlled trials, it has been customary to use clinician-rated scales to demonstrate efficacy, ie, the balance between the specific antidepressant effect and the safety of the drug in terms of adverse drug effects. However, the measure of patient-rated quality of life

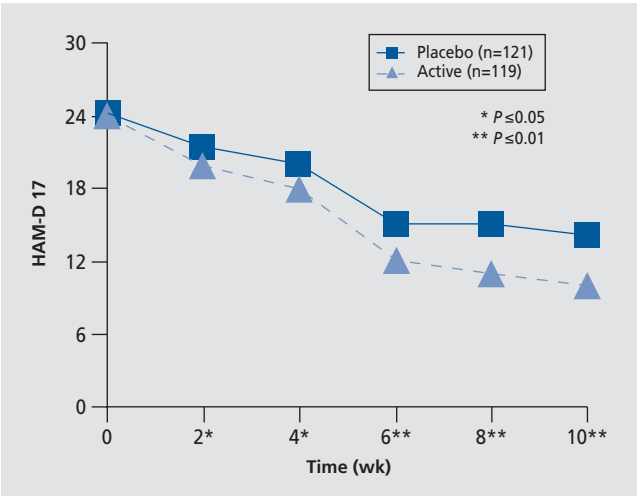

Figure 1. A typical illustration from a placebo-controlled trial with a new potential antidepressant.
Previously presented in a poster at: 4th annual meeting of the Scandinavian College of Neuropsychopharmacology: Elsinore, Denmark; 2005

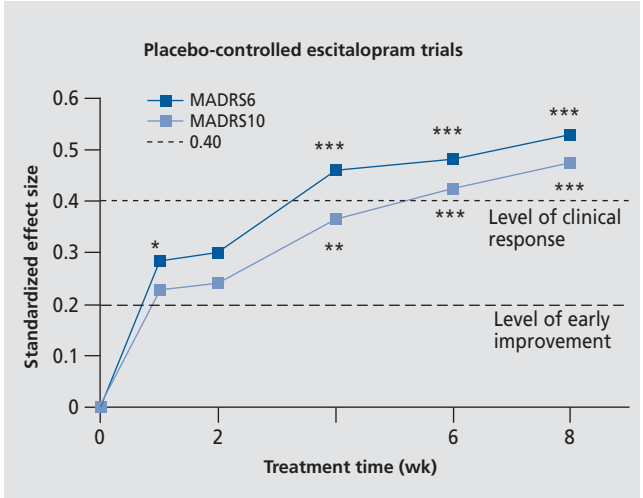

Figure 2. $MADRS_6$ and $MADRS_{10}$ showing the antidepressant activity of 20 mg escitalopram in a placebo-controlled trial in patients with severe major depression.

domains[43] has implied that patient-rated depression rating scales or questionnaires should also be used in placebo-controlled trials. In general, self-rating depression scales such as the Beck Depression Inventory (BDI) or the Zung Depression Scale (SDS) have very rarely been used to demonstrate the clinical effect of SSRIs.[39] Because the classical self-rating scales for depression (BDI, SDS) cover many items, but not all specific items of depression *(Table I)*, it might be appropriate to include the self-rating scale of the HAM-D as released by Bent-Hansen et al.[44] The self-rating version of the HAM-D_6 is shown in *Table III*. Studies are ongoing to evaluate the sensitivity of HAM-D_6 in placebo-controlled trials.

Psychic anxiety (worrying)
Tension (psychic)
Fears
Difficulty in concentration
Muscular tension
Behavior during interview

Table II. The specific items of generalized anxiety in HAM-A_6.

The use of a self-rating version of HAM-D has focused on translation procedures when preparing non-English versions of the scale. This has also implied that the pitfalls of using nonauthorized versions of the HAM-D have been discussed. Even in the most recently published book on assessment scales,[1] the HAM-D_{17} version that is shown is not the original English HAM-D version, although the authors refer to Hamilton's first work with his scale.[45] In the first version of the HAM-D, the item of agitation was measured from 0 to 3, but in the second version, Hamilton changed the scoring to 0-5.[5] The ver-

In this questionnaire you will find six groups of statements. Please choose the one statement in each group that best describes how you have been feeling over the past three days, including today, and mark it with an X in the corresponding box.

(1) During the past three days	
I have been in my usual good mood	0
I have felt a little more sad than usual	1
I have been clearly more sad than usual, but haven't felt helpless or hopeless	2
I have been so gloomy that I briefly have felt overpowered by hopelessness	3
I have been so low in my moods that everything seems dark and hopeless	4
(2) During the past three days	
I have been quite satisfied with myself	0
I have been a little more self-critical than usual with a tendency to feel less worthy than others	1
I have been brooding over my failures in the past	2
I have been plagued with distressing guilt feelings	3
I have been convinced that my current condition is a punishment	4
(3) During the past three days	
My daily activities have been as usual	0
I have been less interested in my usual activities	1
I have felt that I have had difficulty performing my daily activities, but I was still able to perform them with great effort	2
I have had difficulty performing even simple routine activities	3
I have not been able to do any of the most simple day-to-day activities without help	4
(4) During the past three days	
I have felt neither restless nor slowed down	0
I have felt a little slowed down	1
I have felt rather slowed down or have been talking a little less than usual	2
I have felt clearly slowed down or subdued or have talked much less than usual	3
I have hardly been talking at all or felt extremely slowed down all the time	4
(5) During the past three days	
I have been calm and relaxed	0
I have felt a little more tense or insecure than usual	1
I have been clearly more worried or tense than usual, but have not felt that I lost control	2
I have been so tense or worried that I have briefly felt close to panic	3
I have had episodes where I was overwhelmed by panic	4
(6) During the past three days	
I have been as active and have had as much energy as usual	0
I have felt rather low in energy or physically unwell with some bodily pains	1
I have felt very low in energy or had bodily pains	2

Table III. The HAM-D_6 Questionnaire.

sion published by Lam et al[1] is an American version which was not accepted by Hamilton himself,[46] in contrast to the HAM-D_6 version.[47] Hamilton's criticism of the American version included the following: " ... A further deficiency was that it regarded the spontaneous mention of a symptom as indicating greater severity than if it had been elicited by questioning. There are many reasons why patients may not mention a symptom at an interview. For example, they may not think it relevant (eg, feelings of guilt), they may be embarrassed (eg, loss of libido) or they may be too polite to mention to the interviewer that they believe they are suffering from a physical illness ..."

Conclusion

Since the introduction of antidepressants into psychopharmacology in the 1960s, the HAM-D has been the most frequently used rating scale for depresssion. When used as a scale for prediction of outcome with antidepressants, the HAM-D by its total score has obtained limited use analogous to the *DSM-IV* diagnosis of major depression. Among the individual HAM-D items or factors, sleep and agitation are associated with the sedative antidepressants.

Most research has been devoted to the use of HAM-D to discriminate between placebo and active drugs or to show dose-response relationship in patients with major depression. An improvement in the total HAM-D score during a drug trial can, however, not in itself qualify the drug as an antidepressant because the total score is not a sufficient statistic. This implies that the improvement may be found in nonspecific HAM-D factors such as sleep, anxiety, or appetite. To overcome this major pitfall, the specific HAM-D subscales, eg, HAM-D_6 have been discussed with reference also to the analogous subscale from the $MADRS_6$.

The problem of statistical versus clinical significance when analyzing placebo-controlled trials including dose-response relationship has been outlined, with the recommendation to use effect size statistics.

Finally, the pitfall of using unauthorized scale versions has been discussed with reference to self-rating depression scales. ❑

REFERENCES

1. Lam RW, Michalak EE, Swinson RP. *Assessment Scales in Depression, Mania and Anxiety*. London, UK: Taylor and Francis; 2005.
2. Emmelkamp PMG. The additional value of clinimetrics needs to be established rather than assumed. *Psychother Psychosom*. 2004;73:142-144.
3. American Psychiatric Association. *Diagnostic and Statistical Manual Of Mental Disorders*. 4th ed. Washington, DC: American Psychiatric Association; 1994.
4. Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. *Br J Psychiatry*. 1979;134:382-389.
5. Hamilton M. Development of a rating scale for primary depressive illness. *Br J Soc Clin Psychol*. 1967;6:278-296.
6. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. 3rd ed. Washington, DC: American Psychiatric Association; 1980.
7. Bagby RM, Ryder AG, Schuller DR, Marshall MB. The Hamilton Depression Rating Scale: has the gold standard become a lead weight? *Am J Psychiatry*. 2004;161:2163-2177.
8. Bech P, Haaber A, Joyce CRB. Experiments on clinical observation and judgement in the assessment of depression: profiled videotapes and judgement analysis. *Psychol Med*. 1986;16:873-883.
9. Katz MM, Koslow SH, Berman N, et al. A multi-vantaged approach to measurement of behavioral and affect states for clinical and psychobiological research. *Psychol Rev*. 1984;55:619-671.
10. Carney MWP, Roth M, Garside RF. The diagnosis of depressive syndromes and the prediction of ECT response. *Br J Psychiatry*. 1965:111:659-674.
11. Gurney C. Diagnostic scales for affective disorders. *Proceedings of the Fifth World Conference of Psychiatry*. Mexico City, Mexico; 1971:330.
12. Bauer M, Whybrow PC, Angst A, et al. Guidelines for the treatment of unipolar depressive disorders. *World J Psychiatry*. 2002;3:3-43.
13. Posternak MA, Zimmerman M. How accurate are patients in reporting their antidepressant treatment history? *J Affect Disord*. 2003;75:115-124.
14. Faravelli C. Assessment of psychopathology. *Psychother Psychosom*. 2004;73:139-141.
15. Guy W. *Early Clinical Drug Evaluation (ECDEU) Assessment Manual for Psychopharmacology*. Publication No. 76-338. Rockville, Md: National Institute of Mental Health; 1976.
16. Hamilton M. *Methodology of Clinical Research*. Edinburgh, UK: Churchill Livingstone;1974.
17. Bech P, Gram LF, Dein E, Jacobsen O, Vitger J, Bolwig TG. Quantitative rating of depressive states. *Acta Psychiatr Scand*. 1975;51:161-170
18. World Health Organization. *International Classification of Diseases*. 10th revision. Geneva, Switzerland: World Health Organization, 1993.
19. Hamilton M. The effect of treatment on the melancholias (depressions). *Br J Psychiatry*. 1982;140:223-230.
20. Beck AT. *Depression: Clinical, Experimental, and Theoretical aspects*. Philadelphia, Pa: University of Pennsylvania Press; 1967.
21. Bech P. The Bech-Rafaelsen Melancholia Scale (MES) in clinical trials of therapies in depressive disorders: a 20-year review of its use as outcome measure. *Acta Psychiatr Scand*. 2002;106:252-264.
22. Frances A, Pincus HA, Widiger TA, Davis WW, First MB. DSM-IV: work in progress. *Am J Psychiatry*. 1990;147:1439-1448.
23. Bech P, Cialdella P, Haugh M, et al. A meta-analysis of randomised controlled trials of fluoxetine versus placebo and tricyclic antidepressants in the short-term treatment of major depression. *Br J Psychiatry*. 2000;176:421-428
24. Bech P, Tanghøj P, Andersen HF, Overø K. Citalopram dose-response revisited using an alternative psychometric approach to evaluate clinical effects of four fixed citalopram doses compared to placebo in patients with major depression. *Psychopharmacology*. 2002;163:20-25.
25. Bech P, Tanghøj P, Cialdella P, Friis Andersen H, Pedersen AG. Escitalopram dose–response revisited: an alternative psychometric approach to evaluate clinical effects of escitalopram compared to citalopram and placebo in patients with major depression. *Int J Neuropsychopharmacol*. 2004;7:283-290.

Escalas de evaluación en la depresión: limitaciones y obstáculos

*Desde que se introdujeron los antidepresivos en la psicofarmacología, en la década de 1960, la escala de depresión de Hamilton (HAM-D) ha sido la escala de evaluación para la depresión que se ha utilizado con mayor frecuencia. Cuando se emplea la HAM-D como escala para la predicción de la respuesta a los antidepresivos, por su puntuación total, su utilización es limitada tal como ocurre con el diagnóstico de depresión mediante el DSM-IV (*Manual Diagnóstico y Estadístico de los Trastornos Mentales, Cuarta Edición*). Se ha dedicado gran parte de la investigación al uso de la HAM-D para discriminar entre placebo y drogas activas o para mostrar la relación dosis-respuesta en pacientes con depresión mayor. Sin embargo, una mejoría en la puntuación total de la HAM-D durante un ensayo clínico no califica a la droga como antidepresivo, ya que la puntuación total no es un estadístico suficiente. Se comenta el problema de la significación estadística versus la clínica cuando se analizan los ensayos placebo controlados incluyendo la relación dosis-respuesta, con la recomendación de utilizar estadísticos con efecto sobre el tamaño.*

Échelles d'évaluation dans la dépression : pièges et limites

*Depuis l'introduction des antidépresseurs dans la psychopharmacologie au cours des années 1960, l'échelle d'évaluation de la dépression de Hamilton (HAM-D) est celle la plus utilisée pour la dépression. L'HAM-D, lorsqu'elle est utilisée comme échelle de prévision de l'évolution avec les antidépresseurs, atteint ses limites en raison de son score total de même que pour le diagnostic de dépression majeure décrit dans le DSM-IV (*Diagnostic and Statistical Manual of Mental Disorders 4e éd*). Une grande partie des recherches a été consacrée à l'utilisation de la HAM-D pour différencier le placebo du médicament actif ou pour montrer les relations dose-réponse chez les patients ayant une dépression majeure. Une amélioration du score total de la HAM-D au cours d'une étude ne qualifie pas, en soi, un médicament comme antidépresseur, le score total n'étant pas une statistique suffisante. Nous soulignons le problème de la signification statistique versus clinique en analysant des études contrôlées contre placebo qui comprennent des rapports dose-réponse, en recommandant l'utilisation de statistiques qui prennent en compte la taille de l'effet.*

26. Licht RW, Qvitzau S, Allerup P, Bech P. Validation of the Bech-Rafaelsen Melancholia Scale and the Hamilton Depression Scale in patients with major depression: is the total score a valid measure of illness severity? *Acta Psychiatr Scand*. 2005;111:144-149.

27. Bech P. Meta-analysis of placebo-controlled trials with mirtazapine using the core items of the Hamilton Depression Scale as evidence of a pure antidepressive effect in the short-term treatment of major depression. *Int J Neuropsychopharmacol*. 2001;4:337-345.

28. Maier W, Philipp M. Improving the assessment of severity of depressive states: a reduction of the Hamilton Depression Scale. *Pharmacopsychiatry*. 1985;18:114-115.

29. Cleary PJ. Problems of internal consistency and scaling in life event schedules. *J Psychosom Res*. 1981;25:309-320.

30. Detke MJ, Lu Y, Goldstein DJ, McNamara RK, Demitrack MA. Duloxetine 60 mg once daily dosing versus placebo in the acute treatment of major depression. *J Psych Research*. 2002;36:383-390.

31. Ruhe H, Dekker JJ, Peen J, Holman R, de Jonghe F. Clinical use of the Hamilton Depression Scale: is increased efficiency possible? A post hoc comparison of the Hamilton Depression Rating Scale, Maier and Bech subscales, Clinical Global Impression and Symptom Checklist-90 scores. *Comprehens Psychiatry*. 2005;46:417-427.

32. Bech P. Modern psychometrics in clinimetrics: Impact on clinical trials of antidepressants. *Psychother Psychosom*. 2004;73:134-138.

33. Prien RF, Levine J. Research and methodological issues for evaluating the therapeutic effectiveness of antidepressant drugs. *Psychopharmacol Bull*. 1984;20:250-257.

34. Hedges LV, Olkin I. *Statistical Methods for Meta-Analysis*. New York, NY: Academic Press, 1985.

35. Faries D, Herrera J, Rayamajhi J, DeBrota D, Demitrack M, Potter WZ. The responsiveness of the Hamilton Depression Rating Scale. *J Psychiatr Res*. 2000;34:3-10.

36. Bech P. Clinical effects of selective serotonin reuptake inhibitors. In: Dahl SG, Gram LF, eds. *Clinical Pharmacology in Psychiatry*. Springer: Berlin; 1989:81-93.

37. Bech P, Andersen HF. Effective dose of escitalopram in major depressive disorder. *Nord J Pychiatry*. 2005;59:406-407.

38. Entsuah R, Shaffer M, Zhang J. A critical examination of the sensitivity of unidimensional scales derived from the Hamilton Depression Rating Scale of antidepressant drug effects. *J Psychiatr Res*. 2002;36:437-448.

39. Bech P. Pharmacological treatment of depressive disorders: a review. In: Maj, M, Sartorius N, eds. *Depressive Disorders. WPA Series Evidence and Experience in Psychiatry*. 2nd ed. Chichester, UK: John Wiley & Sons; 2002:89-127.

40. Hamilton M. Diagnosis and rating of anxiety. *Br J Psychiatry*. 1969;special issue:76-79.

41. Bech P, Lunde M, Undén M. An inventory for the measurement of generalised anxiety distress symptoms, the GAD-10 Inventory. *Acta Psychiatr Belg*. 2005;105:111-118.

42. Meoni P, Salinas E, Brault Y, Hackett D. Pattern of symptom improvement following treatment with venlafaxine XR in patients with generalized anxiety disorder. *Clin Psychiatry*. 2001;62:888-893.
43. Bech P. *Quality of Life in the Psychiatric Patient*. London, UK; Mosby-Wolfe, 1998.
44. Bent-Hansen J, Lauritzen L, Clemmensen L, Lunde M, Kørner A. A definite and semi-definite questionnaire version of the Hamilton/ Melancholia Scale. *J Affect Disord*. 1995;33:143-150.
45. Hamilton M. A rating scale for depression. *J Neurol Neurosurg Psychiatry*. 1960;23:56-62.
46. Hamilton M, Shapiro CM. Depression. In: Peck DF, Shapiro CM,eds. *Measuring Human Problems*. Chichester, UK: John Wiley; 1990:25-65.
47. Bech P, Kastrup M, Rafaelsen OJ. Mini-compendium of rating scales for states of anxiety, depression, mania, schizophrenia with corresponding DSM-III syndromes. *Acta Psychiatr Scand*. 1986;73 (suppl 326):7-37.