

H.-J. Möller

Methodological aspects in the assessment of severity of depression by the Hamilton Depression Scale

Abstract Observer-rating scales are used for the evaluation of drug trials in depression. One of the most widely used depression rating scales is still the Hamilton Depression Scale (HAMD), which was developed at the time when the first antidepressants were becoming available. Due to its construction it seems to have a specific focus on drug effects of classical antidepressants. As a result of different methodological analyses such as principal component analysis, the Rasch model and facet analysis, a differentiation between core symptoms reflecting the severity of depression and additional symptoms describing other aspects of the symptomatology of depression seems meaningful. The use of the HAMD 6-item score, described primarily by Bech, as the main efficacy criterion in antidepressant drug trials gives a fair estimation of drug-induced changes of severity of depression and avoids bias such as the well-known bias of the HAMD total score in favour of tricyclic antidepressants. This is demonstrated by the evaluation of two sertraline-amitriptyline comparative studies.

Key words Efficacy criteria · antidepressant drug trials · depression rating scales · severity of depression

For Professor Per Bech on the occasion of his 60th birthday. With many thanks for our close academic and personal collaboration.

Introduction

The quantitative evaluation of antidepressant treatment relies on the application of reliable standardised rating scales (Lang et al. 1991). The traditional depression rating scales, such as the Hamilton Depression Scale, were developed in the 1960s, when the first antidepressant drugs were becoming available. They may have been specially designed to measure the efficacy of the classical tricyclic antidepressants (Domken et al. 1994; Hughes et al. 1982; Möller 2000).

There are many ways of classifying scales, but the most important distinguishes between observer-rating scales and self-rating scales (Möller 2000). Both types of scales provide valuable information for the clinician and, in practice, clinical studies sometimes employ both observer- and self-rating scales. Each type of scale is limited in its assessment, providing a complementary view rather than duplication of information. Experts in this field generally agree that observer-rating should be chosen as the principal outcome criterion in antidepressant drug trials. However, self-rating scales can provide additional important information for evaluating therapeutic regimens (Möller and von Zerssen 1995; Möller 1991).

Observer-rating scales have a number of advantages and disadvantages over self-rating scales, and these should be taken into account both in the selection of scales and when reviewing the results of the scales that have been applied. Skilled observers, by means of their clinical training and experience, are able to assess the severity of a patient's symptoms, whereas the patient's point of comparison can only be in relation to his own experience (Hamilton 1976). An observer's ability to assess all grades of severity of illness reliably is affected by his clinical experience and familiarity with the scale he is applying, whereas a severely ill patient may be too ill to complete a self-assessment (Hamilton 1976) or his self-perception may not reflect reality. Likewise, patients with poor literacy skills, deficient vocabulary or diminished concentration span may have problems completing a

Prof. Hans-Jürgen Möller (✉)
Department of Psychiatry
Ludwig-Maximilians-University
Nussbaumstr. 7, 80336 Munich, Germany
Tel.: 089/5160-5501
Fax: 089/5160-5522
E-Mail: hans-juergen.moeller@psy.med.uni-muenchen.de

self-assessment questionnaire (Hamilton 1976). Observers may be subject to a certain bias which can influence the true assessment of a patient's condition. The rating of specific items by an observer can be confounded by general or initial impressions, and the psychiatrist may rate items to fit with that initial impression. The preconception that one therapy is or is not beneficial can also bias the rating of symptoms during the course of treatment, and highlights the need for double-blind clinical trials. Additionally, it may be difficult to differentiate certain personality traits from actual depressive symptoms (Möller 1991; Möller and von Zerssen 1995).

There are several depression observer-rating scales used in psychopharmacological studies. These scales were not designed to be used for diagnostic purposes and are only intended for use once a diagnosis has been made. This type of scale focuses on the psychopathology of depression and is used to measure the severity of symptoms. When used during treatment a scale gives the clinician an assessment of the patient's improved or worsened depressive symptoms. The change in the total score is seen to reflect a change in general severity of depression.

Most often used is still the Hamilton Depression Scale (HAMD) (Hamilton 1960, 1967), although it has been criticised under different aspects. Beside the HAMD, the Montgomery-Asberg Depression Scale is increasingly used in antidepressant drug trials (Möller 2000).

The HAMD was developed to be used on patients who had already been diagnosed as suffering from affective disorders of the depressive type. The scale is used to quantify the results of a clinician's interview with a patient and render the results in a convenient format for statistical analysis. The value of the scale, as with most observer-rating scales, depends on the skill of the interviewer in obtaining the necessary information from his patient during the interview, without using direct or probing questions. The scale has been modified since it was first developed and the original scale is now rarely used. An American version, published in 1976 (Guy 1976), is now the most frequently used. Hamilton's original version contained 17 items, each of which is concerned with a semi-global symptom. These items were chosen because they are the most common symptoms of depressive illness. Some items are defined in terms of a series of categories of increasing intensity (e.g., item 2, guilt), while others are defined by a number of terms with equal values (e.g., item 13, somatic symptoms). However, some symptoms often found with atypical depression are not rated at all (hypersomnia, weight or appetite increase), and some so-called "endogenous" symptoms are not covered, e.g., quality of mood. In this scale, behavioural and somatic features account for at least 50% of the possible total score. An additional four items were originally excluded from the total score (though they were included on the interviewer's form): diurnal variation, because it is neither a measure of depression nor of intensity, but defines the type of depression; and derealisation (or depersonalisation), paranoid symp-

toms and obsessional symptoms, because they occur so infrequently that they do not contribute further information about the majority of patients (Hamilton 1960, 1967; Snaith 1993). The 21-item HAMD (Hamilton 1967) has a total score range from 0 to 56 points. No score on the scale distinguishing normality from morbidity was made by Hamilton; however, the range 0-10 points is generally used to indicate minor or no depression, 10-20 (or, in some trials, 25) major depression and 25 (or 28) more severe depression.

This paper will focus on some methodological aspects of the measurement of the severity of depression by using the Hamilton Depression Scale (HAMD). This topic is of special importance in the context of psychopharmacological studies because the efficacy criteria of drug trials in depression are often related to the total score of the HAMD, which is seen as a measure of the severity of depression.

The structure of severity of depression in the HAMD

Analyses of HAMD data (Bech 1990; Conti and Cassano 1990; Hamilton 1960, 1967; Maier 1990; Maier and Philipp 1985) suggest that a distinction has to be made between those components of depression which are "core symptoms" of severity of major depression and those which reflect different „aspects“ of symptomatology.

Using a probabilistic test model developed by Rasch (1960), Bech et al. (1981) found that the structure or severity of depressive states in the HAMD could be assessed sufficiently by six items (depressed mood, guilt, work and interest, retardation, anxiety psychic, and somatic general). On the basis of these six HAMD items Bech and Rafaelsen (1980) developed their Melancholia Scale (MES). The validity of the subscale of the HAMD consisting of the six core symptoms of depression was partly replicated by Maier and Philipp (1985). The main difference between the two studies is that the item „somatic complaints“ was substituted by the item „agitation“ in the subscale derived by Maier and Philipp (1985).

Among the symptom specific components focussing on different aspects of depressive symptomatology, four dimensions for the 17-item version of HAMD are discussed frequently: somatisation, cognition, retardation and a sleep factor. However, there are considerable discrepancies between the many studies that have been carried out on the HAMD using factor analysis. When Hamilton (1960) published his depression scale, he used factor analysis to delineate the dimensions of severity of depressive states. No general factor was found. In his second study with the HAMD, a general factor of severity was confirmed (Hamilton 1967). As pointed out by Hamilton himself (Hamilton 1986) the reason for this discrepancy between the two studies was that the heterogeneity of the patients regarding severity of depression was larger in the second than in the first study. In general,