Journal of Psychiatric Research 43 (2009) 1049–1056

Contents lists available at ScienceDirect

Journal of Psychiatric Research

journal homepage: www.elsevier.com/locate/jpsychires

Sensitivity of the Montgomery Asberg Depression Rating Scale to response and its consequences for the assessment of efficacy

Gijs Santen [a], Meindert Danhof [a], Oscar Della Pasqua [a,b,*]

[a] Division of Pharmacology, Leiden/Amsterdam Center for Drug Research, P.O. Box 9502, 2300 Leiden, The Netherlands
[b] Clinical Pharmacology and Discovery Medicine, GlaxoSmithKline, Greenford, United Kingdom

ARTICLE INFO

Article history:
Received 26 October 2008
Received in revised form 17 January 2009
Accepted 5 February 2009

Keywords:
Antidepressants
Clinical trials
Endpoint sensitivity
Depression
HAMD
MADRS

ABSTRACT

An increasing body of evidence is available suggesting that the Hamilton Depression Rating Scale (HAMD) is not a sensitive measure of treatment effect. In this investigation, we explore the sensitivity of the individual items of the Montgomery Asberg Depression Rating Scale (MADRS) and compare the consequences of selecting a different scale as primary endpoint in the analysis of efficacy. A graphical approach is proposed for the evaluation of the sensitivity of individual items to response, using data from randomised, placebo-controlled clinical trials in which HAMD and MADRS were measured concurrently. Subsequently, we illustrate the impact of differences in the sensitivity of the primary endpoint for the detection of statistical significance 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, some HAMD subscales still outperform MADRS in the detection of treatment effect. The selection of these subscales as primary endpoints in clinical trials could save over 1/3 in patients compared to the full HAMD whilst keeping the same statistical power.

1. Introduction

There are at least three important reasons to investigate the sensitivity of endpoints used in clinical trials in depression. Firstly, there is little evidence on the contribution of the individual items of the existing clinical scales to the overall assessment of response. Secondly, more than 50% of the performed trials fail, even if efficacious doses of known antidepressants are used (Khan et al., 2002b). Thirdly, the use of more sensitive endpoints may reduce the interferences caused by random noise in response, which enhances the separation of the differences between placebo and active treatment arms, facilitating the detection of statistical significance (Santen et al., 2008). 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 unidimensional measures and objective diagnostic criteria are available (e.g., viral load in HIV infection), the use of rating scales in psychiatric diseases, such as the Hamilton Depression Rating Scale (HAMD) (Hamilton, 1960) 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.

Given the historical evolution, no comprehensive assessment exists of how sensitive these scales are to clinical improvement (i.e., response irrespective of treatment type) and how random noise affects the performance of such a multidimensional summary of symptoms. Most of the efforts in this field have been restricted to criticism on the multidimensionality and unsuitability of the HAMD scale to monitor changes upon treatment (Bech et al., 1980; Moller, 2001), which has led to the 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 et al., 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 Montgomery Asberg Depression Rating Scale (MADRS), which was introduced in 1979. Since then, many studies have used the MADRS as primary endpoint in antidepressant drug trials.

The HAMD and the MADRS are each conducted as a semi-structured observer-rated interview (Table 1). However, the magnitude of item scaling differs between the two instruments. The MADRS

* Corresponding author. Address: Division of Pharmacology, Leiden/Amsterdam Center for Drug Research, P.O. Box 9502, 2300 Leiden, The Netherlands.
E-mail address: odp72514@gsk.com (O.Della Pasqua).

0022-3956/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jpsychires.2009.02.001

Table 1
A list of the individual items of the MADRS and the corresponding HAMD items by symptom domain. The HAMD items included in the subscales 1 and 2 (Santen et al., 2008) and the Bech HAM-D_6 are marked by an 'X'. NA indicates that symptom domain is not applicable to the scale.

Symptom domain	MADRS	HAM-D_{17}	Subscale 1	Subscale 2	Bech HAM-D_6
Mood	Reported sadness	Depressed mood	X	X	X
	Apparent sadness	Suicide	X		
	Inability to feel				
	Suicide				
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		X	
		Middle insomnia			
		Late insomnia		X	
Functional status	Lassitude	Work and interest	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			

has a fixed scaling of seven points (from 0 to 6), whilst the scoring on the HAMD ranges across a smaller number of anchor points, and varies from item to item. To our knowledge, the HAMD, its subscales and the MADRS have not yet been compared simultaneously with respect to their sensitivity to detect treatment effect. Given that the MADRS was especially designed to detect treatment effect, 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 MADRS, HAMD and its subscales as efficacy measure in clinical studies in depression.

In this paper, we apply the same approach used in a previous publication (Santen et al., 2008), in which we have shown that not all items of the HAMD are equally sensitive to response. This methodology allowed us to derive a new subscale (HAM-D_7) as well as to demonstrate its impact on the estimation of the significance level of the separation between placebo and active treatment arms, as compared to the full HAMD scale. In contrast to statistical juggling, 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 clear reduction in the rate 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.

2. Methods

2.1. Clinical 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 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 the HAMD subscales in our previous publication (Santen et al., 2008).

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.

2.2. Sensitivity analysis

Given the multidimensionality of the scales, the sensitivity to clinical response was explored on an item-by-item basis. For that purpose, 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, irrespective of treatment type (placebo or active drug). Even though a definition of response based on the MADRS could be applied, we have chosen to use the HAM-D_{17} as gold standard throughout this investigation to allow consistent comparison between scales. Since clinical response is defined independently of treatment type, the differences between placebo and active treatment were expressed in terms of the fraction of responders in each treatment group.

Dichotomisation into responders *vs.* non-responders and subsequent pooling of the data 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 individual MADRS items across active treatment groups in responders and non-responders. As specified in the study