Available online at www.sciencedirect.com
ScienceDirect
Journal of Psychiatric Research 42 (2008) 1000–1009
JOURNAL OF PSYCHIATRIC RESEARCH
www.elsevier.com/locate/jpsychires

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

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

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

Received 12 April 2007; received in revised form 3 November 2007; accepted 8 November 2007

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 and Philip subscales using a linear mixed-effects modelling approach for repeated measures. The evaluation of two clinical trials revealed endpoint 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 signal-to-noise ratio and could reduce failure rate in efficacy trials with antidepressant drugs.

Keywords: HAMD; MRMM; Depression; Clinical trial design

1. 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.

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

0022-3956/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jpsychires.2007.11.004

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 multi-dimensional (Bech et al., 1980; Moller, 2001). Others have evaluated its sensitivity to drug effect relative to other scales, such as the Montgomery–Asberg depression rating scale (MADRS) (Montgomery et al., 1979; Khan et al., 2004) and the Bech–Rafaelsen Melancholia scale (MES) (Bech and Rafaelsen, 1980). Faries et al. (2000) have shown that a number of published one-dimensional subscales outperform the total HAM-D_{17} 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 and Bech, 2005). The important conclusion from this paper is that the unwanted characteristics of the HAM-D_{17} 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-D_{17} 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-D_{17} 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-D_{17}. Considering that the disease state information in the HAM-D_{17} is unbiased, the objectives of this investigation are to determine the sensitivity of the individual items of the HAM-D_{17} 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.

2. Methods

2.1. 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 Fig. 1.

2.2. 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