Journal of Affective Disorders 225 (2018) 449–452

Contents lists available at ScienceDirect

Journal of Affective Disorders

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

Research paper

Are self-report scales as effective as clinician rating scales in measuring treatment response in routine clinical practice?

Mark Zimmerman*, Emily Walsh, Michael Friedman, Daniela A. Boerescu, Naureen Attiullah

Department of Psychiatry and Human Behavior, Brown Medical School, Rhode Island Hospital, Providence, RI, United States

ARTICLE INFO

Keywords:
Depression
Outcome
Scale
Self-report
Clinician rating
Hamilton
PHQ-9
QIDS
CUDOS

ABSTRACT

Objective: Recent treatment guidelines have suggested that outcome should be measured in routine clinical practice. In the present report from the Rhode Island Methods to Improve Diagnostic Assessment and Services (MIDAS) project, we compared three self-report scales of depressive symptoms and the two most widely used clinician administered scales in treatment studies in their sensitivity to change and evaluation of treatment response in depressed patients treated in routine practice.

Methods: At baseline and 4-month follow-up 153 depressed outpatients with DSM-IV MDD completed the Clinically Useful Depression Outcome Scale (CUDOS), Quick Inventory of Depressive Symptomatology—Self-report version (QIDS-SR), and Patient Health Questionnaire (PHQ-9). The patients were rated on the 17-item Hamilton Depression Rating Scale (HAMD) and the Montgomery-Asberg Depression Rating Scale (MADRS). On each scale treatment response was defined as a 50% or greater reduction in scores from baseline.

Results: While there were some differences in the percentage of patients considered to be responders on the different scales, a large effect size was found for each scale, with little variability amongst the scales. The level of agreement between the three self-report scales and the clinician rating scales was approximately the same

Limitations: The present study was conducted in a single clinical practice in which the majority of the patients were white, female, and had health insurance.

Discussion: When measuring outcome in clinical practice the magnitude of change in depressive symptoms is as great on self-report scales as on clinician rating scales.

1. Introduction

In psychiatry, quantified assessments of outcome are not the standard of care. Instead, in mental health clinical settings outcome evaluations are typically based on unstructured interactions that yield *unquantified* judgments of progress. This is at variance with other areas of medical care in which outcome is determined, in part, on the change of a numerical value. Body temperature, blood pressure, cholesterol values, blood sugar levels, cardiac ejection fraction, thyroid stimulating hormone levels, and white blood cell counts are examples of quantifiable variables that are used to evaluate treatment progress. Quantifiable outcome measures exist for most major psychiatric disorders, yet they are rarely used in routine clinical practice (Gilbody et al., 2002; Zimmerman and McGlinchey, 2008).

The quantitative measurement of treatment outcome has long been an integral component of research investigations of the efficacy and effectiveness of care. Recently, some investigators and treatment guidelines have suggested that measurement tools should be used to monitor the course of treatment in clinical practice (American Psychiatric Association, 2010; Harding et al., 2011; National Collaborating Centre, 2009; Trivedi et al., 2006). A better understanding of the effectiveness of psychiatric treatment in clinical practice depends, in part, on systematically measuring outcome. To accomplish this, reliable, valid, informative, and user-friendly scales are necessary. Clinicians are already overburdened with paperwork, and adding to this load by suggesting repeated detailed evaluations with such instruments as the Hamilton Rating Scale for Depression (HAMD) (Hamilton, 1960) or the Montgomery Asberg Depression Rating Scale (MADRS) (Montgomery and Asberg, 1979) is unlikely to meet with success. Clinician-rated scales are time consuming, require training to ensure the ratings are reliable and valid, and may be prone to clinician bias. Self-report questionnaires are inexpensive in terms of professional time needed for incorporation into the clinical encounter, they do not require special training for administration, and they correlate highly with clinician ratings. With modern technology, computer administered self-report assessments enable the conduct of large-scale outcome studies in

* Correspondence to: 146 West River Street, Providence, RI 02904, United States.
E-mail address: mzimmerman@lifespan.org (M. Zimmerman).

http://dx.doi.org/10.1016/j.jad.2017.08.024
Received 21 December 2016; Received in revised form 31 May 2017; Accepted 10 August 2017
Available online 15 August 2017

clinical practice at low cost (Zimmerman and Martinez, 2012). Moreover, self-report scales are free of clinician bias, and are therefore free from the potential risk of clinician overestimation of patient improvement (which might occur when there is incentive to document treatment success).

A meta-analysis of treatment studies of depression found that effect sizes of treatment as assessed by self-administered scales were smaller than the effect sizes as assessed by clinician-rated measures (Cuijpers et al., 2010). Little research has compared the effect sizes of self-report and clinician rated scales in routine clinical practice. While many self-report scales have been developed to measure the severity of depression (Nezu et al., 2000) Zimmerman et al. (2008b), in discussing the use of self-report scales in routine clinical practice, recommended measures that assess the DSM-IV criteria for major depressive disorder (MDD) that are available for clinical use at no cost. Several such scales exist (Bech et al., 2001; Kroenke et al., 2001; Rush et al., 2003, 1996; Zimmerman et al., 2008a, 2004). In consideration of increasing calls to demonstrate the effectiveness of treatment in routine practice, and the lower clinical burden imposed by self-report scales compared to clinician-rated scales, it is important to determine if the method of assessing outcome will significantly influence conclusions about the degree of treatment effectiveness.

Accordingly, in the present report from the Rhode Island Methods to Improve Diagnostic Assessment and Services (MIDAS) project, we compared three self-report scales assessing the DSM-IV symptom criteria for MDD and the 2 most widely used clinician administered scales in their sensitivity to change and evaluation of treatment response in depressed patients treated in routine practice.

2. Methods

One hundred fifty-three patients diagnosed with DSM-IV MDD who presented for treatment to the Rhode Island Hospital Department of Psychiatry outpatient practice (n = 78), or who were in ongoing treatment and had their medication changed due to lack of efficacy (n = 75), were evaluated at baseline and at 4-month follow-up. The mean interval between the baseline and follow-up evaluations was 16.4 weeks (SD = 4.2 weeks). Not all available patients participated in the study due to the lack of availability of raters or the treating psychiatrist did not refer the patient to the study. Approximately half of the patients were diagnosed with MDD based on the Structured Clinical Interview for DSM-IV (SCID) (First et al., 1995), whereas the other patients were diagnosed on the basis of an unstructured clinical interview. The sample included 42 (27.5%) men and 111 (72.5%) women who ranged in age from 18 to 79 years (M = 43.7, SD = 13.6). The Rhode Island Hospital institutional review committee approved the research protocol, and all patients provided informed, written consent.

The patients completed the CUDOS, PHQ-9, and QIDS at baseline and follow-up and were evaluated with the 17-item HAMD, and MADRS blind to the completion of the self-report scales.

The CUDOS contains items assessing all of the DSM-IV inclusion criteria for MDD (Zimmerman et al., 2008a). The respondent is instructed to rate the symptom items on a 5-point Likert scale indicating "how well the item describes you during the past week, including today" (0 = not at all true/0 days; 1 = rarely true/1–2 days; 2 = sometimes true/3–4 days; 3 = usually true/5–6 days; 4 = almost always true/every day). Compound DSM-IV symptom criteria referring to more than one construct (e.g. problems concentrating or making decisions; insomnia or hypersomnia) were subdivided into their respective components and a CUDOS item was written for each component. Total scores range from 0 to 64.

Similar to the CUDOS, the QIDS uses 16 items to assess the DSM-IV symptom criteria (Rush et al., 2006). However, the format of the 2 questionnaires differs. On the QIDS each symptom is assessed by a group of 4 statements, and the respondent selects the item that best describes how they have been feeling. Not every item contributes to the total score. In scoring the QIDS the highest score is used of the 4 items assessing sleep disturbance (initial, middle or terminal insomnia, or hypersomnia), the 2 items assessing psychomotor disturbance (agitation, retardation), and the 4 items assessing appetite and weight disturbance. Total scores on the scale range from 0 to 27.

The PHQ-9 contains 9 items corresponding to the DSM-IV major depressive disorder criteria (Kroenke et al., 2001). Unlike the CUDOS and QIDS, the PHQ-9 assesses compound symptom criteria with a single item. For example, the PHQ-9 assesses insomnia and hypersomnia, and reduced or increased appetite, with a single item. The respondent is instructed to rate the symptom items on a 4-point Likert scale indicating how often they have been bothered by the symptom over the past 2 weeks (0 = not at all; 1 = several days; 2 = more than half the days; 3 = nearly every day). Total scores on the scale range from 0 to 27.

The HAMD is the most commonly used clinician-rated outcome scale in depression treatment studies (Zimmerman et al., 2015). The original rating form included 21 items, though Hamilton (1960) indicated that only the first 17 items should contribute to the total scale score because one of the last four items represented depressive type rather than depression severity (diurnal mood variation), and three other items did not occur with sufficient frequency (derealization, paranoia, and obsessional symptoms). Nine of the 17 items are rated from 0 to 4 whereas 8 items are rated 0–2, thus the maximum score is 52. We examined the total scale score as well as a 6-item subscale representing the core symptoms of depression that has been found to be more sensitive to change (Bech, 2001; Bech et al., 2010).

The MADRS is the second most commonly used clinician-rated scale, and has been used to evaluate outcome in antidepressant efficacy trials with increasing frequency in recent years (Zimmerman et al., 2015). Whereas the HAMD was intended as a measure of the severity of depressive symptoms, the MADRS was designed to be particularly sensitive to change in patients treated with antidepressant medication. The MADRS consists of 10 items rated from 0 to 6, thus the maximum score is 60.

2.1. Data analyses

Pearson correlations were computed between the change in scores on each of the measures. For each scale, we used paired *t*-tests to compare follow-up scores to baseline values. We computed the effect size (Cohen's d) on each of the measures. An effect size of .2 was considered small, .5 medium, and .8 large (Cohen, 1988). We used McNemar's test to compare the percentage of patients classified as being treatment responders on each measure. Treatment response was defined as a 50% or greater reduction in scores from baseline. The kappa statistic was used to determine the level of agreement between the scales in identifying treatment response. The data was analyzed using SPSS version 22.0.

3. Results

There was no difference in the amount of change in the patients who presented for treatment versus those in ongoing treatment who had their medication changed therefore the data from these 2 groups was combined. On each scale, the patients showed significant levels of improvement from baseline to follow-up (Table 1). A large effect size was found for each scale (Table 1), with little variability amongst the scales.

All correlations between the scales in change in scores from baseline to 4 months were significant (Table 2). The correlations amongst the self-report scales (mean r = .76) and amongst the clinician rated scales (mean r = .81) were higher than the correlations between the self-report and clinician rated scales (.69).

The data in Table 3 shows the number of patients considered to be treatment responders according to the different scales. Significantly more patients were considered to be a treatment responder on the MADRS than each of the other scales except the PHQ-9 ($HAMD_6$,