Journal of Clinical Epidemiology 77 (2016) 4–6

Journal of Clinical Epidemiology

COMMENTARIES

Are more responsive depression scales really superior depression scales?

Eiko I. Fried*

Faculty of Psychology and Educational Sciences, University of Leuven, Tiensestraat 102, Leuven 3000, Belgium

Accepted 23 May 2016; Published online 28 May 2016

1. Introduction

In contemporary clinical practice and research, many different instruments are used to measure depression severity. Santor et al. [1] identified no fewer than 280 scales developed in the last century, many of which are still in use. The responsiveness of these various depression scales to treatment—and the comparison of responsiveness among scales—has become an important research topic. Responsiveness is defined as a scale's ability to detect clinical change and to measure the true effect of treatment on depression severity [2]. There are numerous ways to assess and compare responsiveness. In this issue, Kounali et al. [3] introduce a meta-analytic method that allows to compare the responsiveness of different scales by constructing a network of randomized trials (i.e., scales can be compared across different data sets). In their proof of principle analysis, the authors examine instruments including the Patient Health Questionnaire (PHQ-9) [4] and the Beck Depression Inventory (BDI) [5], concluding that some scales perform better than others.

This is consistent with the literature: responsiveness is often treated as a psychometric property [3,6], and scales more responsive to treatment have been advanced as superior scales that should be used in clinical trials [3,7,8]. Bech [8], for instance, has called for the use of the six-item subscale of the Hamilton Rating Scale for Depression (HRSD-6) instead of the 17- and 21-item versions because, among other reasons, "fewer patients are then needed to identify antidepressant effect in controlled trials" (p. 310) (see also [9]). Similarly, Kounali et al. [3] advance that "responsiveness [...] can help guide the choice of outcome measures in clinical trials".

This line of reasoning, that has crucial implications for future research, relies on a strong assumption: that the most responsive depression scales provide valid measures of patients' improvement—that they adequately assess clinical change. We elucidate below why this is not the case and why scale responsiveness alone should not be treated as inherently desirable.

2. What does responsiveness measure

Faced with the challenge to assess whether patients are doing better over the course of treatment, depression severity is commonly operationalized as a sum of depression symptoms (rating scales add up all symptoms to a total score), and the decrease of this sum score is understood as a measurement of patients' improvement. The responsiveness of a scale is its ability to detect this clinical change. Imagine we enroll 200 patients with an average of 19 points on the BDI at baseline and an average of 13 points after 8 weeks of treatment. Consistent with the literature, we interpret the reduction of 6 points as the BDIs responsiveness, which we can standardize to compare it to the responsiveness of other scales. But do these 6 points adequately reflect clinical change? As we will see below, this conceptualization rests on three implicit assumptions of unidimensionality, temporal invariance, and content validity that are unlikely to hold in depression scales.

2.1. Lack of unidimensionality and temporal measurement invariance

Our understanding of the sum of BDI symptoms as measurement of one underlying condition (depression) requires the BDI to be unidimensional. The same holds for interpreting the change in total scores over time as measurement of patients' depression improvement. Unidimensionality means that all BDI symptoms such as crying, loss of interest in sex, agitation, and feelings of guilt should measure the same underlying disorder depression. In psychometric terms, we expect that a single factor accounts for (nearly) all covariance among symptoms so that symptoms are conditionally independent when controlling for the factor. Half a century of psychometric literature contradicts the notion of unidimensionality [10–12]: virtually all depression scales are

Funding: This study is supported in part the Research Foundation Flanders (G.0806.13), the Belgian Federal Science Policy within the framework of the Interuniversity Attraction Poles program (IAP/P7/06), and the grant GOA/15/003 from University of Leuven.

Conflicts of interest: None.

* Corresponding author. Tel.: (+32) 16 379525.

E-mail address: eiko.fried@gmail.com

http://dx.doi.org/10.1016/j.jclinepi.2016.05.004

multifactorial in depressed samples, including shorter instruments like the PHQ-9 [6]. (Note that unidimensionality and higher levels of internal consistency have been reported in healthy or mixed populations—often samples at exit timepoints of clinical trials [10]—but not for depressed populations.) But if unidimensionality does not hold—if the BDI measures a set of (somewhat related) constructs—what does a responsiveness of 6 points mean exactly?

Second, our interpretation of a 6-point change as actual improvement of the patients' depression requires temporal measurement invariance (MI): that the relationship between depression symptoms and the underlying construct that is measured (depression) does not change over time [13]. If MI holds, the 19 points at baseline have the same psychometric meaning as the 13 points 8 weeks later, and changes in the BDI sum score over time represent actual differences in the (set of) construct(s) we track. There are varying degrees to which MI can be violated, and depression scales assessed in clinical trials consistently fail to meet the most basic level of MI: the number of factors required to describe the covariance among symptoms changes systematically over time [10]. This means that our observed difference of 6 points does not adequately reflect changes of the underlying (set of) construct(s) the BDI supposedly measures and offers limited or even misleading insights into patients' progress.

Taken together, there is nothing wrong with the mere mathematical operation of subtracting 13 from 19 that results in 6. However, advancing this difference of 6 points as valid measurement of clinical change—which is exactly how responsiveness is defined (e.g., [6])—seems hardly defensible because the difference of sum scores describes multidimensional constructs that change over time.

2.2. *Content validity*

A pragmatist may try to salvage the situation by suggesting the BDI sum (and its change) to be a mere index score of problems (a formative latent variable) instead of a measurement of depression (a reflective latent variable) [10,14,15]. In the case of responsiveness, such a solution would be short lived because the lack of content validity of depression rating scales amounts to another main challenge for interpreting more responsive scales as psychometrically desirable scales.

Depression instruments differ substantially regarding their symptom content, which is the main reason that scales show different rates of responsiveness. The BDI, for example, captures primarily cognitive symptoms like pessimism and worthlessness, whereas the HRSD focuses on somatic symptoms like sexual dysfunction, and the Center for Epidemiological Studies Scale includes frequent crying, talking less, and perceiving others as unfriendly. None of these symptoms appear in criteria for an episode of major depression (MD) as listed in the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) [16], which are in turn captured by the PHQ-9. Considering these pronounced differences, it is not surprising that different depression scales measure different constructs [10,12,17], and various studies have shown that choosing a specific depression scale over another one can substantially change results of empirical investigations [1,18,19,20].

This raises the question whether short scales like the PHQ-9 or the HRSD-6 that have been suggested to be superior scales based on their responsiveness to treatment [3,6,8,9] possess adequate content validity [21]—whether they capture all relevant facets of depression. Responsiveness is defined as detecting clinical change, which requires to measure most clinically relevant problems patients exhibit. Considering the heterogeneity of the depressive syndrome, it is unlikely that short scales can easily fulfill this requirement. Concentration problems, for instance, have been shown to be among the most debilitating and central depression symptoms [14,22], but are not captured by the HRSD-6. The PHQ-9, in turn, does not assess anxiety or irritability that are not only common and debilitating problems among depressed patients, reduce remission rates and prolong remission, but are also markers of a more severe, chronic, and complex depressive illness [23–25].

It has also been suggested that shorter scales like the HRSD-6 are more valid measures of clinical change than longer ones [9] because antidepressant treatment often adversely impacts on depression symptoms such as sleep and weight changes, sexual dysfunction, and suicidal ideation. I agree that the full HRSD is indeed sensitive to picking up side effects of antidepressant treatment [26]—but if antidepressant treatment substantially exacerbates existing problems of depressed patients, using a scale that ignores these problems contradicts the very definition of responsiveness as detecting clinical change. Imagine we query our 200 patients about 30 relevant problems at enrollment and 8 weeks later simply remove all items from our new scale on which patients did not improve; this would result in a highly responsive scale. But this instrument would not capture patients' clinical improvement, it would merely conceal problems depressed individuals still struggle with after the clinical trial ends. In addition, our scale would be biased by the type of treatment: we would get a rather different scale after 8 weeks of treatment with atypical antidepressants compared to 8 weeks of cognitive behavioral therapy.

A final example elucidates that responsiveness depends on the assessed symptoms. In a recent study, Arnow et al. [27] investigated the response of patients with melancholic, atypical, and anxious depression to three antidepressant drugs and found no differences among groups regarding treatment response. They queried patients via the Quick Inventory of Depressive Symptoms 16 to examine treatment response that only captures the 9 DSM-5 criterion symptoms of MD, while patients were diagnosed with subtypes of MD defined by problems that go beyond these nine symptoms, including despair for melancholia and paralysis for atypical depression. While

DSM symptoms of MD certainly play an important role for the well-being of patients exhibiting specific depression subtypes, they are not sufficient to adequately capture severity or treatment response, and an analysis of a broader range of symptoms—for instance, via the Inventory of Depressive Symptomatology 30—may have altered study results significantly.

3. Conclusion

We set out to explore whether responsiveness of depression scales adequately captures patients' depression improvement—which is necessary for the routine interpretation of more responsive scales to be psychometrically superior scales. This is unlikely to be the case for a majority of scales as three key requirements are violated: unidimensionality, temporal MI, and adequate content validity. Responsiveness is moderated by the particular symptoms instruments encompass, and short scales in particular are prone to ignore relevant aspects of the clinical syndrome. This may explain why patients' recovery of impaired functioning often lags behind about half a year after symptomatic remission [28].

Multivariate analyses of scales that capture a broad range of depression symptoms may provide better insights into how patients are doing in clinical trials [26,29].

Acknowledgments

The author would like to express his sincerest thanks to Aurélie Wagener, Susann Ulrich, and Rogier Kievit for comments on a prior version of the manuscript and the Editor for inviting this manuscript.

References

[1] Santor DA, Gregus M, Welch A. Eight decades of measurement in depression. Measurement 2009;4:135–55.
[2] Wright JG, Young L. A comparison of different indices of responsiveness. J Clin Epidemiol 1997;50:239–46.
[3] Kounali D, Button K, Lewis G, Ades AE. The relative responsiveness of test instruments can be estimated using a meta-analytic approach: an illustration with treatments for depression. J Clin Epidemiol 2016, http://dx.doi.org/10.1016/j.jclinepi.2016.03.005, [Epub ahead of print].
[4] Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001;16:606–13.
[5] Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J. An inventory for measuring depression. Arch Gen Psychiatry 1961;4:561–71.
[6] Titov N, Dear BF, Mcmillan D, Anderson T, Zou J, Sunderland M. Psychometric comparison of the PHQ-9 and BDI-II for measuring response during treatment of depression. Cogn Behav Ther 2011;40:126–36.
[7] Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. Br J Psychiatry 1979;134:382–9.
[8] Bech P. The responsiveness of the different versions of the Hamilton Depression Scale. World Psychiatry 2015;14:309–10.
[9] Bech P. Is the antidepressive effect of second-generation antidepressants a myth? Psychol Med 2010;40:181–6.
[10] Fried EI, van Borkulo CD, Epskamp S, Schoevers RA, Tuerlinckx F, Borsboom D. Measuring depression over time… or not? Lack of unidimensionality and longitudinal measurement invariance in four common rating scales of depression. Psychol Assess 2016;. http://dx.doi.org/10.1037/pas0000275. [Epub ahead of print].
[11] Gullion CM, Rush AJ. Toward a generalizable model of symptoms in major depressive disorder. Biol Psychiatry 1998;44:959–72.
[12] Shafer AB. Meta-analysis of the factor structures of four depression questionnaires: Beck, CES-D, Hamilton, and Zung. J Clin Psychol 2006;62:123–46.
[13] Meredith W. Measurement invariance, factor analysis and factorial invariance. Psychometrika 1993;58:525–43.
[14] Fried EI, Epskamp S, Nesse RM, Tuerlinckx F, Borsboom D. What are "good" depression symptoms? Comparing the centrality of DSM and non-DSM symptoms of depression in a network analysis. J Affect Disord 2015;189:314–20.
[15] Bollen KA, Lennox R. Conventional wisdom on measurement: a structural equation perspective. Psychol Bull 1991;110:305–14.
[16] APA. Diagnostic and statistical manual of mental disorders. 5th ed. Washington, DC: American Psychiatric Association; 2013.
[17] van Loo HM, de Jonge P, Romeijn JW, Kessler RC, Schoevers RA. Data-driven subtypes of major depressive disorder: a systematic review. BMC Med 2012;10:156.
[18] Zimmerman M, Martinez JH, Friedman M, Boerescu D, Attiullah N, Toba C. How can we use depression severity to guide treatment selection when measures of depression categorize patients differently? J Clin Psychiatry 2012;73:1287–91.
[19] Snaith P. What do depression rating scales measure? Br J Psychiatry 1993;163:293–8.
[20] Polaino A, Senra C. Measurement of depression: comparison between self-reports and clinical assessments of depressed outpatients. J Psychopathol Behav Assess 1991;13:313–24.
[21] Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychol Bull 1955;52:281–302.
[22] Fried EI, Nesse RM. The impact of individual depressive symptoms on impairment of psychosocial functioning. PLoS One 2014;9:e90311.
[23] Judd LL, Schettler PJ, Coryell W, Akiskal HS, Fiedorowicz JG. Overt irritability/anger in unipolar major depressive episodes: past and current characteristics and implications for long-term course. JAMA Psychiatry 2013;70:1171–80.
[24] Fava M, Rush AJ, Alpert JE, Balasubramani GK, Wisniewski SR, Carmin CN, et al. Difference in treatment outcome in outpatients with anxious versus nonanxious depression: a STAR*D report. Am J Psychiatry 2008;165:342–51.
[25] ten Have M, Lamers F, Wardenaar K, Beekman A, de Jonge P, van Dorsselaer S, et al. The identification of symptom-based subtypes of depression: a nationally representative cohort study. J Affect Disord 2016;190:395–406.
[26] Fried EI, Nesse RM. Depression sum-scores don't add up: why analyzing specific depression symptoms is essential. BMC Med 2015;13:1–11.
[27] Arnow BA, Blasey C, Williams LM, Palmer DM, Rekshan W, Schatzberg AF, et al. Depression subtypes in predicting antidepressant response: a report from the iSPOT-D trial. Am J Psychiatry 2015;172:743–50.
[28] McKnight PE, Kashdan TB. The importance of functional impairment to mental health outcomes: a case for reassessing our goals in depression treatment research. Clin Psychol Rev 2009;29:243–59.
[29] Hieronymus F, Emilsson JF, Nilsson S, Eriksson E. Consistent superiority of selective serotonin reuptake inhibitors over placebo in reducing depressed mood in patients with major depression. Mol Psychiatry 2016;21:523–30.