Journal of Affective Disorders 127 (2010) 122–129

Contents lists available at ScienceDirect

Journal of Affective Disorders

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

Research report

Defining successful treatment outcome in depression using the PHQ-9: A comparison of methods

Dean McMillan [a,*], Simon Gilbody [a], David Richards [b]

[a] Hull York Medical School and Department of Health Sciences, University of York, United Kingdom
[b] Mood Disorders Centre, School of Psychology, University of Exeter, United Kingdom

ARTICLE INFO

Article history:
Received 21 December 2009
Received in revised form 27 April 2010
Accepted 28 April 2010
Available online 31 May 2010

Keywords:
Clinical significance
Depression
PHQ-9

ABSTRACT

Background: Although the PHQ-9 is widely used in primary care, little is known about its performance in quantifying improvement. The original validation study of the PHQ-9 defined clinically significant change as a post-treatment score of ≤9 combined with improvement of 50%, but it is unclear how this relates to other theoretically informed methods of defining successful outcome. We compared a range of definitions of clinically significant change (original definition, asymptomatic criterion, reliable and clinically significant change criteria a, b and c) in a clinical trial of a community-level depression intervention.
Method: Randomised Control Trial of collaborative care for depression. Levels of agreement were calculated between the standard definition, other definitions, and gold-standard diagnostic interview.
Results: The standard definition showed good agreement (kappa>0.60) with the other definitions and had moderate, though acceptable, agreement with the diagnostic interview (kappa = 0.58). The standard definition corresponded closely to reliable and clinically significant change criterion c, the recommended method of quantifying improvement when clinical and non-clinical distributions overlap.
Limitations: The absence of follow-up data meant that an asymptomatic criterion rather than remission or recovery criteria were used.
Conclusion: The close agreement between the standard definition and reliable and clinically significant change criterion c provides some support for the standard definition of improvement. However, it may be preferable to use a reliable change index rather than 50% improvement. Remission status, based on the asymptomatic range and a lower PHQ-9 score, may provide a useful additional category of clinical change.

1. Introduction

The PHQ-9 (Kroenke et al., 2001) is a widely used self-report measure of depression that is brief, easy to administer and has well-established psychometric properties (Lee et al., 2007). For these reasons it has been recommended as an integral part of the management of depression in primary care, including tracking symptom change and defining successful treatment outcome to inform treatment decisions (Clark et al., 2009; Dejesus et al., 2007). However, little is known about the performance of the measure in quantifying clinically significant improvement.

A small number of studies have established that the measure if sensitive to change (Cameron et al., 2008; Lowe et al., 2004a) and one study has used minimal clinically important difference (MCID) criteria to define a change of 5-points or more as indicating MCID (Lowe et al., 2004b). Apart from the study by Lowe et al. (2004b), the only other guidance on defining clinically significant change comes from

* Corresponding author. Department of Health Sciences, University of York, YO10 5DD, United Kingdom. Tel.: +44 1904 321359; fax: +44 1904 321388.
E-mail address: dm645@york.ac.uk (D. McMillan).

0165-0327/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jad.2010.04.030

the original validation study of the PHQ-9, which recommended a score of ≥10 to indicate the presence of probable depression (Kroenke et al., 2001) based on an analysis of sensitivity and specificity data. In that dataset, scores of ≥10 indicated an increased probability of receiving a diagnosis of major depressive disorder (MDD); whereas few people who scored ≤9 met diagnostic criteria for MDD. On the basis of this, Kroenke et al. (2001) made a recommendation that a post-treatment score of ≤9 along with the commonly used criterion of a 50% reduction in scores could be used to define clinically significant improvement. Kroenke et al. (2001), however, pointed out that their definition of improvement was provisional and that further work was needed to validate it.

There are a number of alternative methods of conceptualising improvement on measures of psychological functioning in general and depression in particular with clear theoretical underpinnings, and it is not clear how the standard definition recommended in the original validation study relates to these. These include recovery and remission criteria (Frank et al., 1991) and the concepts of reliable and clinically significant change (Jacobson and Truax, 1991).

Frank et al. (1991) provided several conceptual definitions of improvement in depression that have proved influential in current thinking about defining treatment outcome (Keller, 2003) The concept of remission requires a period, typically at least several weeks, in which a person remains in an asymptomatic range, defined as no or very few symptoms. Recovery requires that the person remains in this asymptomatic range, but for a longer duration. The concepts of an asymptomatic range, remission and recovery may be important in measuring treatment outcome for depression. Rates of relapse and recurrence following successful treatment for depression remain high. A consistent finding is that people who are classed as improved but continue to have some residual symptoms have a substantially higher rate of relapse and recurrence than those who meet criteria for remission or recovery (Paykel et al., 1995).

Reliable and clinically significance change criteria (Jacobson and Truax, 1991) are among the most commonly used method of quantifying improvement in studies of psychological treatments (Ogles et al., 2001) and have been recommended as a standard reporting strategy for all published research involving these types of interventions (Evans et al., 1998). At a conceptual level, clinically significant change defines improvement as a move from a clinical to a non-clinical range. Jacobson and Truax (1991) provide several operational definitions of cut-off points to distinguish between these ranges based on the central location and distribution of scores for a clinical and non-clinical group. As an additional criterion, the change in scores must be greater than that which could be due to the inherent unreliability of the measure.

It is not clear how the standard definition of improvement for the PHQ-9 relates to other commonly used methods of defining improvement on psychological measures. The aim of this study is to examine the performance of this standard definition by comparing it to other commonly used definitions of improvement. As an additional index of corroboration, we compared the agreement between these definitions and a gold-standard diagnostic interview.

2. Method

2.1. Sample

The sample was taken from a randomised control trial of collaborative care for depression (Richards et al., 2007). Participants were recruited from primary care services, and were included if they were aged above 18 years, had received a diagnosis of depression by a GP, and scored ≥5 on the Structure Clinical Interview for DSM-IV defined major depressive disorder (MDD) (Spitzer et al., 1992). Exclusion criteria included active suicidal plans, primary drug or alcohol dependence and some types of depression (post-natal, bereavement-related, depression with a physical cause). 114 participants were recruited. The majority of the sample were female (77%) and the mean age of the sample was 43.3 years (sd = 13.6; range 18–77). Approximately half of the participants were married (54%) and a similar proportion was employed (49%).

2.2. Measures

The PHQ-9 is a nine-item measure of depression based on the Diagnostic and Statistical Manual (DSM) diagnostic criteria for major depressive disorder (Kroenke et al., 2001). Each item is rated on a 0 to 3 scale relating to the frequency of depressive symptoms (0 = "not at all" to 3 = "nearly every day"). Scores range from 0 to 27 with higher scores indicating a greater severity of depression. The original validation report on the PHQ-9 indicated adequate psychometric properties (Kroenke et al., 2001).

The Structured Clinical Interview for DSM-IV (SCID) is a semi-structured interview for making DSM axis I diagnoses (Spitzer et al., 1992), and is extensively used as a research instrument. Trained research assistants conducted all of the SCID interviews.

2.3. Procedure

Patients with depression managed in primary care practices were randomised to case management ($n = 41$) or usual care conditions ($n = 73$; 38 individually randomised, 35 randomised by primary care cluster). The PHQ-9 was completed at pre-treatment and 3-months post-randomisation. The SCID was conducted at pre-treatment to establish eligibility for inclusion and repeated at 3-months post-randomisation. See Richards et al. (2007) for a detailed description of trial methods.

2.4. Analysis

We defined treatment response in five ways.

2.4.1. Standard definition

The standard cut-off point requires a person to move from a depressed range, defined as a score of ≥10, pre-treatment to a non-depressed range, defined as a score of ≤9, post-treatment. In addition, the person's score had to improve by 50% or more from pre- to post-treatment.

2.4.2. Asymptomatic range

There are no accepted definitions of an asymptomatic range, remission or recovery for the PHQ-9. Frank et al. (1991) provided operational definitions of an asymptomatic range for a number of standardised depression instruments. On the Beck Depression Inventory (BDI; Beck et al., 1961), the asymptomatic range was defined as a score of ≤8. This closely matches the demarcation of the non-depressed range (0–9) and mild-moderate range (10–18) recommended by the originators of the test (Beck et al., 1988). For the 17-item Hamilton Rating Scale for Depression (Hamilton, 1960), a score of ≤7 was taken to indicate an asymptomatic range, which again is close to the demarcation of the non-depressed (0–6) and mild depression range (7–17) on the measure (Dozois and Dobson, 2002). The original validation study of the PHQ-9 recommends that scores of 0–4 are in the minimal range and score of 5–9 the mild depression range. We therefore defined the asymptomatic range as a score of ≤4. The absence of follow-up data in the current study meant that were we were not able to include a duration criteria to define full remission or recovery.

2.4.3. Reliable and clinically significant change

Jacobson and Truax (1991) provide three strategies for defining clinically significant change. Criterion a defines the non-clinical range as scores more than two standard deviations below the mean of a clinical sample. Criterion b identifies the non-clinical range as a score within two standard deviations of the non-clinical mean. Criterion c uses the score at which the probability of coming from a clinical and non-clinical distribution is equal; scores below this point are classified as the non-clinical range. Clinically significant change for each of these criteria requires that a person is above the cut-off pre-treatment (i.e. is in the clinical range) but below it at post-treatment. We used pre-treatment data ($n=114$) from the current study to derive the clinical mean (17.3) and standard deviation (5.0). We used data from the original validation study (Kroenke et al., 2001) to provide the non-clinical mean (3.3) and standard deviation (3.8).

The reliable change index identifies the level of change (pre-treatment to post-treatment) that is required on a measure for the change to be classified as reliable. The reliable change index uses the standard error of the difference, which indicates the distribution of change scores that would be observed were no change to have taken place, to identify the level of change that would be unlikely to be observed ($p<0.05$) were no change to have taken place (Jacobson and Truax, 1991). If the difference between pre-treatment and post-treatment scores for a person exceeds this level, the person is classified as making reliable change. This reliable change index can be used to identify people showing a reliable improvement or a reliable deterioration in scores. The calculation requires data on the standard deviation of a pre-treatment clinical group and an indication of the reliability of the measure. The internal reliability estimate (Cronbach's alpha) of 0.89 for the PHQ-9 reported in the original validation study was selected to make this calculation (Kroenke et al., 2001) along with the pre-treatment standard deviation from the current study.

On the basis of these data, reliable and clinically significant change criteria were calculated, and these are summarised in Table 1 along with improvement criteria for the other definitions.

Table 1
Operational definitions of improvement.

Definition of improvement	Pre-treatment score must be:	Post-treatment score must be:	Improvement in score must be:
Standard definition	≥10	≤9	≥50% of pre-treat score
Asymptomatic	≥5	≤4	N/A
RCSC (criterion a) [1]	≥8	≤7	≥5
RCSC (criterion b) [2]	≥11	≤10	≥5
RCSC (criterion c) [3]	≥10	≤9	≥5

[1] RCSC (criterion a): Reliable and clinically significant change using criterion a (reliable improvement and a score 2 sd below the clinical mean).
[2] RCSC (criterion b): Reliable and clinically significant change using criterion b (reliable improvement and a score within 2 sd of the non-clinical mean).
[3] RCSC (criterion c): Reliable and clinically significant change using criterion c (reliable improvement and a greater likelihood of the person being in the non-clinical than the clinical distribution).

2.4.4. Statistical methods

Absolute rates of improvement were calculated for each definition of improvement. Kappa was used to estimate the level of agreement between the different methods and with a post-treatment diagnosis of major depressive disorder based on the SCID; ranges for kappa (e.g., 0.61–0.80 = good agreement) are based on the recommendations of Altman (1991). Treatment effect sizes (measured as Odds Ratios with 95% confidence intervals) were calculated between improvement status and treatment group for each definition of improvement. We adjusted confidence intervals for clustering within practices using the Huber–White estimator (Ukoumunne et al., 1999; White, 1980).

3. Results

The standard definition suggested a similar level of improvement (36.5%) to the other definitions, with the exception of the asymptomatic criterion (27.1%), which suggested lower rates of improvement than all other definitions (Table 2). Of those participants who scored in

Table 2
Proportion showing improvement using different definitions.

Definition of improvement	Meets criteria for improvement % (*n*)	Does not meet criteria for improvement % (*n*)	Pre-treatment score in non-clinical range % (*n*)
Standard cut-off point	36.5 (35)	54.2 (52)	9.4 (9)
Asymptomatic (≤4)	27.1 (26)	72.9 (70)	0 (0)
RCSC (criterion a) [1]	36.5 (35)	61.5 (59)	2.1 (2)
RCSC (criterion b) [2]	40.6 (39)	46.9 (45)	12.5 (12)
RCSC (criterion c) [3]	40.6 (39)	50.0 (48)	9.4 (9)

[1] RCSC (criterion a): Reliable and clinically significant change using criterion a (reliable improvement and a score 2 sd below the clinical mean).
[2] RCSC (criterion b): Reliable and clinically significant change using criterion b (reliable improvement and a score within 2 sd of the non-clinical mean).
[3] RCSC (criterion c): Reliable and clinically significant change using criterion c (reliable improvement and a greater likelihood of the person being in the non-clinical than the clinical distribution).

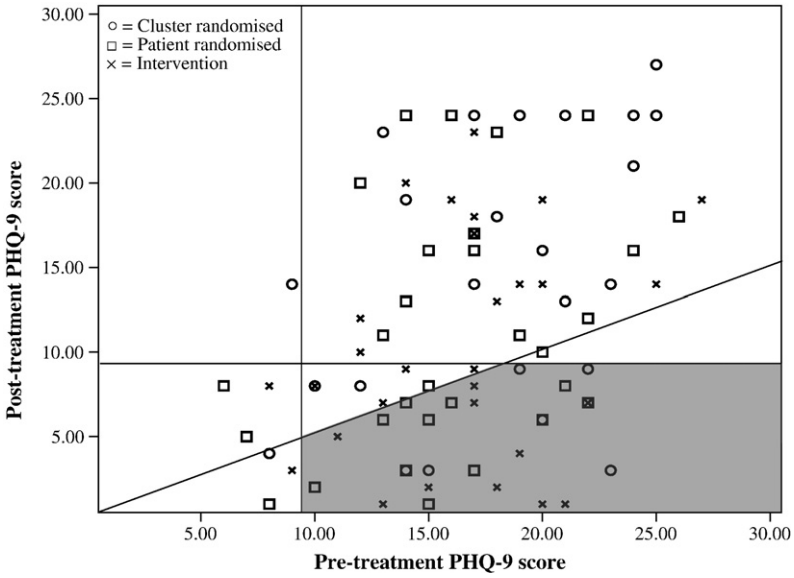

Fig. 1. Standard definition (≤9 and 50% improvement).

the clinical range pre-treatment for all definitions ($n = 84$), approximately half (53.8%) did not meet improvement criteria for any definition and 23.8% met criteria for all definitions; there were disagreements between the definitions for 22.6% of participants.

Figs. 1 to 3 plot pre-treatment scores against post-treatment scores for the sample and indicate improvement criteria for three of the definitions (original definition, asymptomatic criterion, reliable and clinically significant change criterion c). Scores to the right of the vertical line indicate those people scoring above the minimum pre-treatment criterion for the definition; those scores below the horizontal line indicate those people scoring below the post-treatment cut-off point for the definition. In Fig. 1 the diagonal line indicates the point of 50% improvement. Fig. 3 uses the "tramline" display recommended by Jacobson and Truax (1991) to illustrate reliable change. The central diagonal indicates no change. Scores below the lower diagonal exceed the minimum level of improvement for reliable improvement, in this case five points. Scores above the upper diagonal indicate those people showing reliable deterioration in their scores from pre- to post-treatment (deterioration in score of five points or more). In all of the figures, the shaded areas indicate the combinations of pre- and post-treatment scores that would meet the improvement criteria for the particular definition.

Table 3 summarises kappa levels of agreement between the definitions and gold-standard diagnostic interview. The standard definition of improvement showed a very good level

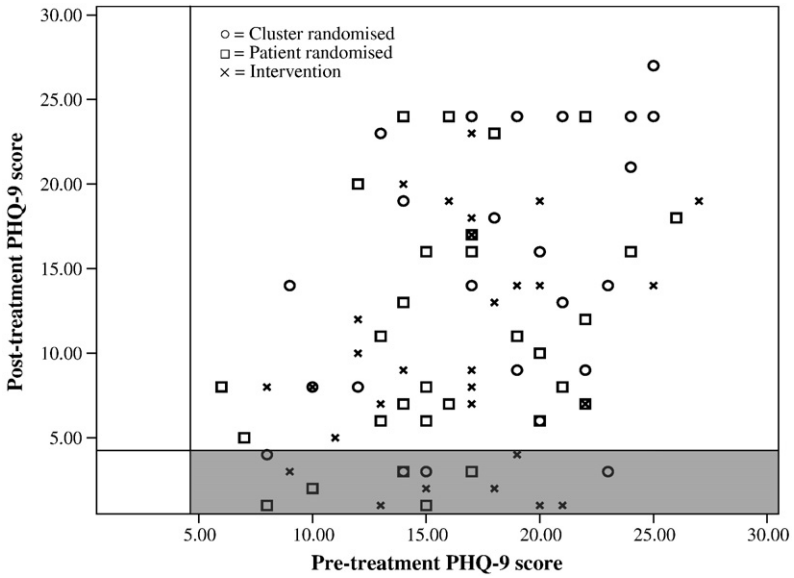

Fig. 2. Asymptomatic range (≤4).

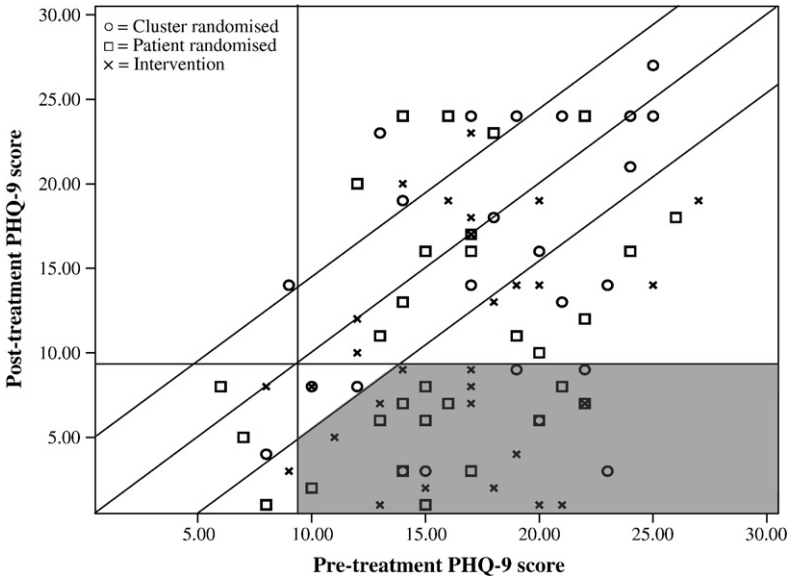

Fig. 3. Reliable and clinically significant change criterion c.

of agreement (kappa>0.80) with reliable and clinically significant change criterion c as well as the other Jacobson and Truax (1991) methods of quantifying change. It also showed a good level of agreement (kappa>0.60) with the asymptomatic criterion. Agreement with the gold-standard diagnostic interview was moderate (kappa = 0.58), though still within the acceptable range. The standard definition tended to be more conservative than the SCID rating. For 17 out of the 18 cases in which a disagreement occurred between the two, the SCID rated the person as improved when the standard definition did not.

Table 4 summarises the proportion in each treatment group showing improvement and Fig. 4 summarises the odds ratios adjusted for clustering within practices. The odds ratio and 95% CI for the standard definition were similar to those of the reliable and clinically significant change criteria b and c,

Table 3
Agreement (kappa) between definitions and structured clinical interview.

Definition of improvement	Asymptomatic (≤4)	RCSC (criterion a) [1]	RCSC (criterion b) [2]	RCSC (criterion c) [3]	SCID diagnosis
Standard cut-off point (≤9)	0.64	0.88	0.88	0.91	0.58
Asymptomatic (≤4)	–	0.69	0.53	0.56	0.36
RCSC (criterion a) [1]	–	–	0.81	0.84	0.54
RCSC (criterion b) [2]	–	–	–	0.98	0.67
RCSC (criterion c) [3]	–	–	–	–	0.62

Note: All kappa values are significant at $p<0.001$.
[1] RCSC (criterion a): Reliable and clinically significant change using criterion a (reliable improvement and a score 2 sd below the clinical mean).
[2] RCSC (criterion b): Reliable and clinically significant change using criterion b (reliable improvement and a score within 2 sd of the non-clinical mean).
[3] RCSC (criterion c): Reliable and clinically significant change using criterion c (reliable improvement and a greater likelihood of the person being in the non-clinical than the clinical distribution).

Table 4
Proportion showing improvement by different treatment group.

Definition of improvement	Proportion of intervention showing improvement % (*n*/total *n*)	Proportion of patient randomised showing improvement % (*n*/total *n*)	Proportion of cluster randomised showing improvement % (*n*/total *n*)
Standard cut-off point (≤9)	43.8 (14/32)	45.2 (14/31)	29.2 (7/24)
Asymptomatic (≤4)	34.3 (12/35)	23.5 (8/34)	22.2 (6/27)
RCSC (criterion a) [1]	45.7 (16/35)	43.8 (14/32)	18.5 (5/27)
RCSC (criterion b) [2]	54.8 (17/31)	50.0 (15/30)	30.4 (7/23)
RCSC (criterion c) [3]	53.1 (17/32)	48.4 (15/31)	29.2 (7/24)

[1] RCSC (criterion a): Reliable and clinically significant change using criterion a (reliable improvement and a score 2 sd below the clinical mean).
[2] RCSC (criterion b): Reliable and clinically significant change using criterion b (reliable improvement and a score within 2 sd of the non-clinical mean).
[3] RCSC (criterion c): Reliable and clinically significant change using criterion c (reliable improvement and a greater likelihood of the person being in the non-clinical than the clinical distribution.

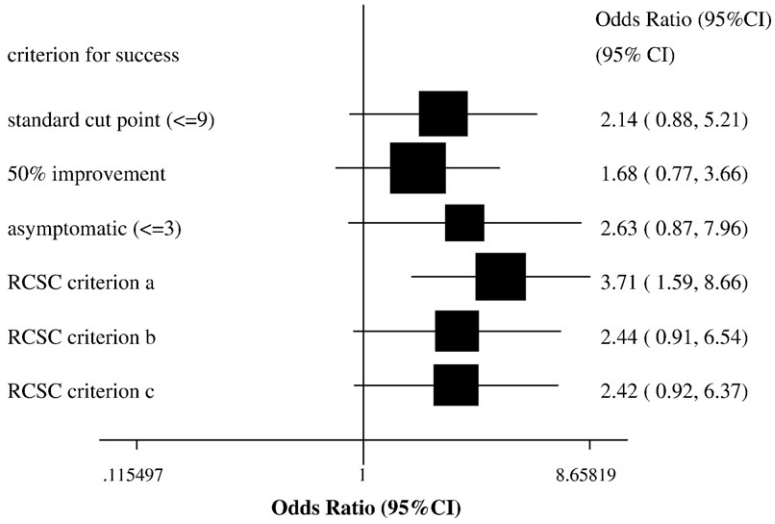

Fig. 4. Odds ratios and 95% confidence intervals for treatment group and improvement status adjusted for clustering.

and all crossed 1. Only the 95% CI for the reliable and clinically significant change criterion a did not include 1.

4. Discussion

If the PHQ-9 is to be of use in clinical practice it will be necessary to define clinically significant improvement on the measure. Although Kroenke et al. (2001) offered such a definition, they were careful to point out that it was provisional and further work was needed to validate it. To this end, this study compared the definition of Kroenke et al. (2001) with other theoretically informed methods of quantifying improvement as well as a gold-standard diagnostic interview.

The standard definition showed a good or very good level of agreement with the other definitions and a moderate level of agreement with the gold-standard diagnostic interview. It is of note that the reliable and clinically significant change criterion c used the same cut-off point of $\leq$9. Of the three methods of defining clinically significant change described by Jacobson and Truax (1991), criterion c is recommended as the most appropriate strategy when the clinical and non-clinical distributions overlap as they frequently do for psychological measures, including the PHQ-9. This provides some corroboration for the standard definition.

The two methods, however, use different additional criteria for the minimal level of improvement needed between pre- and post-treatment (minimum of five points improvement vs. minimum of 50% improvement). Despite this, in practice they classified similar people as improved or not improved (see Figs. 1 and 3). The five-point criterion has the advantage of recognising the inherent unreliability of psychological measures, and specifies a minimum level of change that must be exceeded before someone is classified as showing improvement. This helps reduce the likelihood of falsely classifying someone as making improvement when in fact the change in scores may be due to the unreliability of the measure. However, the 50% criterion will also require this level of improvement when combined with the criterion that a person's pre-treatment score must be $\geq$10, because anyone scoring 10 or more pre-treatment and improving by 50% or more will also improve by at least five points.

While it may not make much practical difference which additional criterion is used alongside the cut-off point of $\leq$9, there may be a number of reasons to prefer the reliable change index over 50% improvement. First, the method has a theoretical basis whereas the 50% improvement criterion has been criticised as arbitrary (Evans et al., 1998). Secondly, the level of improvement it requires corresponds with the level of improvement specified by the minimal clinically important difference criterion (Lowe et al., 2004b), which should permit comparisons across studies using these different criteria. Thirdly, as further data on the means, standard deviations and reliability of the measure accumulate it will be possible to further refine the definition of improvement, which may include developing separate criteria for different types of clinical groups.

Perhaps the most important argument for the use of the reliable change index, however, is that it allows clinicians to detect a reliable deterioration in symptoms, which on the basis of the current psychometric data would be a deterioration of five points or more. While variations in scores during treatment of less than five points either way of the initial starting score suggest that no change has taken place, if a score increases by five points or more it indicates a worsening of depressive symptoms. These two situations may require different clinical management strategies.

The rate of improvement based on an estimate of the asymptomatic range differed from those of the other definitions and also showed the lowest rate of agreement with the SCID. The criterion was by far the most stringent of the definitions in that it requires a score similar to or below the mean of a non-depressed group. If a person is to be classified as improved on the basis of this criterion, he or she needs to move from a clinical range to a point lower than a substantial proportion of people who are not depressed.

The differences between the asymptomatic range and the other definitions reflect different conceptual approaches to improvement. While the other definitions are attempting to estimate a move from a clinical to a non-clinical range, an

asymptomatic range along with the additional duration criteria for remission and recovery, aim to index a level of improvement that substantially lowers the probability of relapse or recurrence. A residual level of symptoms, and therefore an increased probability of relapse, may be within a non-clinical range. While a score in the range of, for example, 5 to 9 may indicate that a person's mood is comparable to someone who is not depressed, a score in this range may mean something different in terms of the probability of a future episode of depression for a person who has never experienced depression than for someone who has recently recovered from an episode.

A remission criterion could, therefore, act as an additional tier of improvement. A score of ≤9 combined with an improvement of five points or more would indicate clinically significant improvement; a score that meets remission criteria would demarcate an additional level of improvement indicating a markedly reduced risk of relapse and recurrence. Whether or not a score of ≤4 represents an adequate definition of an asymptomatic range remains to be seen. Ideally, prospective data on the relationship between post-treatment scores on the PHQ-9 and future depressive episodes would be needed to establish this cut-off point.

The recommendations for quantifying improvement on the PHQ-9 are provisional for a number of reasons. The data are taken from a small randomised trial (Richards et al., 2007), and the study used a number of exclusion criteria. It is unclear whether similar results would have been found in a larger, more representative clinical sample. Although the trained research assistant conducted the SCID interviews, no data are available on reliability. The recommended cut-off points for reliable and clinically significant criteria can vary depending on the psychometric data used in their calculation. It is possible that different results would have been obtained were different psychometric values used and these would not have shown agreement with the standard definition. It may be helpful, therefore, to use systematic review strategies to increase the reliability of estimates of the psychometric values and to explore the reasons for variation in these values between studies. It is also important to note that the results apply only to quantifying improvement for major depressive disorder. Different cut-off points may be needed for other types of depressive difficulties or in situations in which the PHQ-9 is used as an index of general psychological functioning.

The standard definition of improvement on the PHQ-9 showed generally good agreement with other theoretically informed methods of quantifying improvement. In particular, it used the same cut-off point as Jacobson and Truax (1991) clinically significant change criterion c, the recommended method of quantifying improvement when there is an overlap in the distribution of clinical and non-clinical scores. This provides some support for defining clinically significant change as requiring a move from a score of ≥10 pre-treatment to a score of ≤9 post-treatment. Although there were differences in the minimal degree of change required by the standard definition and that of criterion c, when combined with the additional criteria (≥10 pre-treatment, ≤9 post-treatment) they will broadly agree in their categorisation of people as improved or not improved. However, there may be some reasons to favour the reliable change index over the standard definition, which requires an improvement of 50% or more. This suggests that in addition to a score moving from ≥10 pre-treatment to ≤9 post-treatment, the improvement must be ≥5 points on the measure. The use of the reliable change index also allows the clinician and researcher to identify those people showing a reliable deterioration in depressive symptoms, which can be defined as a deterioration of ≥5 points. This may be important for treatment planning. Finally, it may be useful to establish remission and recovery criteria for the PHQ-9, which are likely to require a cut-off point substantially lower than nine to indicate an asymptomatic range.

Role of funding source

The randomised trial was funded by MRC grant no. G03000677; ID: 68073, International Standard RCT no.: ISRCT63222059. The researchers worked independently of the research funder in designing of the study, the collection, analysis and interpretation of data, in the writing of the report, and in the decision to submit the paper for publication.

Conflict of interest

All authors declare that they have no conflict of interest.

Acknowledgements

Thank you to Dr Peter Bower for comments on an earlier version of this manuscript.

References

Altman, D., 1991. Practical Statistics for Medical Research. Chapman & Hall, London.

Beck, A.T., Steer, R.A., Carbin, M.G., 1988. Psychometric properties of the Beck Depression Inventory: twenty-five years of evaluation. Clinical Psychology Review 8, 77–100.

Beck, A.T., Ward, C.H., Mendelson, M., Mock, J., Erbaugh, J., 1961. An inventory for measuring depression. Archives of General Psychiatry 4, 561–571.

Cameron, I.M., Crawford, J.R., Lawton, K., Reid, I.C., 2008. Psychometric comparison of PHQ-9 and HADS for measuring depression severity in primary care. British Journal of General Practice 58, 32–36.

Clark, D., Layard, R., Smithies, R., Richards, D., Suckling, R., Wright, B., 2009. Improving access to psychological therapy: initial evaluation of two UK demonstration sites. Behaviour Research and Therapy 47, 910–920.

Dejesus, R.S., Vickers, K.S., Melin, G.J., Williams, M.D., 2007. A system-based approach to depression management in primary care using the Patient Health Questionnaire-9. Mayo Clinic Proceedings 82, 1395–1402.

Dozois, D.J.A., Dobson, K.S., 2002. Depression. In: Anthony, M.M., Barlow, D.A. (Eds.), Handbook of Assessment and Treatment Planning for Psychological Disorders. Guilford Press, New York, pp. 259–299.

Evans, C., Margison, F., Barkham, M., 1998. The contribution of reliable and clinically significant change methods to evidence-based mental health. Evidence-based Mental Health 1, 70–72.

Frank, E., Prien, R.F., Jarrett, R.B., Keller, M.B., Kupfer, D.J., Lavori, P.W., Rush, A.J., Weissman, M.M., 1991. Conceptualization and rationale for consensus definitions of terms in major depressive disorder. Remission, recovery, relapse, and recurrence. Archives of General Psychiatry 48, 851–855.

Hamilton, M., 1960. A rating scale for depression. Journal of Neurology, Neurosurgery and Psychiatry 23, 56–62.

Jacobson, N., Truax, P., 1991. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. Journal of Consulting and Clinical Psychology 59, 12–19.

Keller, M.B., 2003. Past, present, and future directions for defining optimal treatment outcome in depression: remission and beyond. Journal of the American Medical Association 289, 3152–3160.

Kroenke, K., Spitzer, R.L., Williams, J.B., 2001. The PHQ-9: validity of a brief depression severity measure. Journal of General Internal Medicine 16, 606–613.

Lee, P.W., Schulberg, H.C., Raue, P., Kroenke, K., 2007. Concordance between the PHQ-9 and the HSCL-20 in depressed primary care patients. Journal of Affective Disorders 99, 139–145.

Lowe, B., Kroenke, K., Herzog, W., Grafe, K., 2004a. Measuring depression outcome with a brief self-report instrument: sensitivity to change of the Patient Health Questionnaire (PHQ-9). Journal of Affective Disorders 81, 61–66.

Lowe, B., Unutzer, J., Callahan, C.M., Perkins, A.J., Kroenke, K., 2004b. Monitoring depression treatment outcomes with the patient health questionnaire-9. Medical Care 42, 1194–1201.

Ogles, B.M., Lunnen, K.M., Bonesteel, K., 2001. Clinical significance, history, application, and current practice. Clinical Psychology Review 21, 421–446.

Paykel, E.S., Ramana, R., Cooper, Z., Hayhurst, H., Kerr, J., Barocka, A., 1995. Residual symptoms after partial remission: an important outcome in depression. Psychological Medicine 25, 1171–1180.

Richards, D.A., Lovell, K., Gilbody, S., Gask, L., Torgerson, D., Barkham, M., Bland, M., Bower, P., Lankshear, A.J., Simpson, A., 2007. Collaborative care for depression in UK primary care: a randomized controlled trial. Psychological Medicine 38, 279–287.

Spitzer, R.L., Williams, J.B., Gibbon, M., First, M.B., 1992. The Structured Clinical Interview for DSM-III-R (SCID). I: history, rationale, and description. Archives of General Psychiatry 49, 624–629.

Ukoumunne, O.C., Gulliford, M.C., Chinn, S., Sterne, J.A.C., Burney, P.G.J., Donner, A., 1999. Methods in health service research: evaluation of health interventions at area and organisation level. British Medical Journal 319, 376–379.

White, H., 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 48, 817–838.