Title

Cut points on the Patient Health Questionnaire (PHQ-9) that predict response to cognitive–behavioral treatments for depression

Journal

General Hospital Psychiatry, 37(5)

ISSN

0163-8343

Authors

Schueller, Stephen M
Kwasny, Mary J
Dear, Blake F
et al.

Publication Date

2015-09-01

DOI

10.1016/j.genhosppsych.2015.05.009

Peer reviewed

HHS Public Access

Author manuscript

Gen Hosp Psychiatry. Author manuscript; available in PMC 2016 September 01.

Published in final edited form as:

Gen Hosp Psychiatry. 2015 ; 37(5): 470–475. doi:10.1016/j.genhosppsych.2015.05.009.

Cut Points on the Patient Health Questionnaire (PHQ-9) that Predict Response to Cognitive-Behavioral Treatments for Depression

Stephen M. Schueller, PhD[a], **Mary J. Kwasny, ScD**[a], **Blake F. Dear, PhD**[b], **Nickolai Titov, PhD**[b], and **David C. Mohr, PhD**[a]

Stephen M. Schueller: Schueller@northwestern.edu; Mary J. Kwasny: m-kwasny@northwestern.edu; Blake F. Dear: blake.dear@mq.edu.au; Nickolai Titov: nick.titov@mq.edu.au; David C. Mohr: d-mohr@northwestern.edu

[a]Northwestern University, Feinberg School of Medicine, Department of Preventive Medicine, 750 N. Lake Shore Drive, 10th Floor, Chicago, IL 60611, United States

[b]Macquarie University, Department of Psychology, New South Wales 2109, Australia

Abstract

Objective—Monitoring depressive symptoms during treatment can guide clinical decision-making and improve outcomes. The aim of this study was to explore values on the Patient Health Questionnaire (PHQ-9) that could predict response to treatment.

Method—Data came from two independent trials, including three treatment modalities of cognitive-behavioral treatment for depression. Four hundred eighty-seven participants who either met DSM-IV criteria for major depressive disorder or had PHQ-9 scores consistent with a diagnosis of depression were included in our analyses. Participants either received 18 weeks of telephone or face-to-face (n = 279), or 8 weeks of web-delivered (n = 208) cognitive-behavioral therapy. Depressive symptoms, evaluated using the PHQ-9, were reported every 4 weeks in the telephone and face-to-face trial and weekly in the web-delivered intervention trial.

Results—Optimal cut points for predicting end of treatment response were consistent in both trials. Our results suggested using cut points of a PHQ-9 ≥17 at Week 4, and PHQ-9 ≥13 at Week 9 and PHQ-9 ≥ 9 at Week 14.

Conclusions—Consistent specified cut points were found within trials included. These cut points may be valuable for algorithms to support clinical decision-making.

Keywords

depression; cognitive-behavioral therapy; treatment; computer/Internet technology; measurement

Correspondence to: Stephen M. Schueller, Schueller@northwestern.edu.

1. Introduction

Depression is a common psychiatric disorder with an estimated prevalence of between 6.6–10.3% in the general population [1]. With such a common problem and in light of recent changes in healthcare policy, including the requirements of the Affordable Care Act and the Mental Health Parity Act, depression treatment requires efficient models of care. Furthermore, healthcare practices increasingly require the monitoring of symptoms and functional outcomes to allocate and measure the quality of these services [2]. The most commonly administered measure of depressive symptoms is the Patient Health Questionnaire (PHQ-9) [3]. The PHQ-9 has demonstrated usefulness as a screening method and for monitoring response to treatment overtime [4,5]. Although guidelines for the PHQ-9 exist, including when to treatment should be initiated or when remission is achieved [4, 6], more research is required to know how to gauge levels of symptoms during treatment and whether one is certain levels of symptoms are suggestive of likelihood of ultimate treatment response in a given modality.

As such, monitoring symptoms during the course of treatment is a critical step towards improved clinical decision-making. Effective treatments for depression exist, although successful reduction of symptoms is not universal [7]. Guidelines recommend continuing treatment until full remission is achieved [6], yet only about half of all patients who receive an appropriate length of an evidence-based psychotherapy or an adequate dose of pharmacotherapy reach remission [8,9]. In order to improve outcomes and cost-effectiveness of treatments, three general approaches have been investigated including adjusting the course of a single form of treatment, sequencing treatments, and stepped care. The level of symptoms that should be used to guide treatment in this regard, however, remains an open question.

Adjusting the course of treatment, whether it be psychotherapy or pharmacotherapy, can substantially increase the likelihood that a given individual will experience reduction in his or her depressive symptoms. In psychotherapy, when clinicians receive feedback corresponding to symptoms early during treatment, it can decrease the proportion of patients who reach the end of treatment without benefiting [10,11]. Similarly, many patients who do not respond to an initial antidepressant may respond to some other medication or to a combination of medications [12, 13, 14]. In light of this, it is critical to monitor symptoms during the course of treatment. This allows the provider to gain knowledge if treatment is progressing as expecting and to make corrections if necessary to improve the likelihood of eventual treatment response.

A growing focus has examined sequencing or adding treatment modalities. Collaborative care, one of the most researched of these methods, involves a team of health care professionals, often including primary care physicians, psychiatrists, nurses, psychologists, and social workers, who can provide a range of treatments including pharmacotherapies and psychotherapies [15, 16]. A structured managed plan is used to provide the ‘right’ care at the ‘right’ time. Patients are monitored and adjustments to care are based on change in symptom severity. Stepped care has emerged as a model in response to growing concerns about managing the prevalence of mental health problems in light of the cost of treatment [17]. In

stepped care, patients begin with a low intensity treatment. Progress is monitored systematically, and patients who do not respond adequately are stepped up to more intensive treatments [18]. A systematic review suggests that stepped care approaches can be effective in reducing depressive symptoms, yet revealed a wide variety of criteria used to step patients to next level treatments [19]. Of note, several of the studies used cut points on measures of depressive symptoms as the criterion. However, those cut points varied widely across studies, ranging from values based on overall norms for the scale (for example scores of 1 standard deviation above the mean on the CES-D, translating to scores greater than or equal to 16) or other clinical intuitions or guidelines (PHQ-9 greater than or equal to 10, PHQ-9 greater than or equal to 5). Thus, empirical investigations are needed to determine the proper cut points.

One study investigated the use of cut points in psychotherapy (consisting of a combination of cognitive-behavioral and interpersonal psychotherapy principles) and psychotherapy plus medication [20]. However, this study used change scores (percent change on symptom measures), rather than absolute scores, which may be less useful for clinical practice, as clinicians are much more likely to look at symptom severity scores rather than calculate the change over time. Thus, we were interested in investigating optimal cut points using absolute scores at different time points as the predictor of ultimate treatment response.

In the current article, we explore optimal cut points in two trials that contain three types of treatments: face-to-face and telephone psychotherapy, and a low intensity web-delivered treatment. These three treatment methods represent a range of therapeutic modalities currently used in behavioral care. Primary outcomes for these trials are reported elsewhere [21, 22].

2. Method

2.1. Study Design and Participants

Data for these analyses came from two published trials. One was a trial comparing face-to-face cognitive behavioral therapy (CBT) to telephone CBT for the treatment of major depressive disorder [21]. The other trial examined a well-validated web-delivered intervention [22]. From the first trial, a total of 279 patients who had baseline and end of treatment PHQ-9 scores available were included. The mean age at the start of this trial was 48 (*SD* = 13) and ranged from 19 to 86; 77.1% were female, and 32.6% were on antidepressant medication. The second trial compared transdiagnostic versus disorder specific web-delivered CBT. The first trial required a diagnosis of major depressive disorder according to the *Diagnostic and Statistical Manual of Mental Disorders,* 4th edition [23] for inclusion in the trial, whereas, the second did not as it included a transdiagnostic CBT treatment. As our interest was investigating clinical cut points for depressive symptoms and to make the comparison groups more equitable we restricted our analysis of participants from the second trial to those with a PHQ-9 ≥10 at baseline (all patients in the first trial met this criterion). In the second trial, a total of 208 patients met this criterion. The mean age at the start of that trial was 43, with standard deviation (*SD*) of 12 (ranging from 18–63); 72.6% were female, and 39.9% were on antidepressant medication. This resulted in a total sample of 487 patients across all analyses. Although, an appropriate length of CBT is

typically 16 to 20 weeks, web-based treatment is typically shorter (e.g., 6–10 weeks), as that seems to be the limit that people stay engaged. Each of the trials used appropriate lengths for the corresponding treatment modality. The Human Subjects Committee of the Institutional Review Boards of the corresponding institutions approved each trial.

2.2. Description of the Interventions

2.2.1. CBT study: Telephone Therapy (T-CBT) and Face-to-Face Therapy (FtF-CBT)—Detailed protocol information is reported in the paper addressing the primary outcomes [21]. In short, both treatments followed the same CBT protocol with the only difference being delivery medium. The protocol had been previously validated for telephone administration [24, 25]. Participants received 18 45-minute sessions: 2 sessions weekly for the first 2 weeks, followed by 12 weekly sessions, with 2 final booster sessions during 4 weeks. Participants received a CBT workbook consisting of 8 chapters. Core lessons included behavioral activation, cognitive restructuring, and social support and optional lessons covered common comorbidities, including anxiety and worry, relaxation training, communication and assertiveness training, anger management, and insomnia. Nine PhD-level psychologists provided the treatment. Because post-treatment depression outcomes were equivalent for FtF-CBT and T-CBT [21], primary analyses grouped these two treatments.

2.2.2. I-CBT study: Web-delivered CBT—Participants were randomly allocated to receive access to either a self-guided or therapist-guided version of either the Depression Course or the Wellbeing Course. The structure and content for these two treatments were similar, however the Depression Course specifically targeted symptoms of depression, while the Wellbeing Course targeted symptoms of depression and common anxiety disorders [22]. Both treatments were based on a CBT model and contained five online lessons delivered over eight weeks, a homework assignment for each lesson, regular automated reminder and notification emails, and additional written resources about skills helpful to people with depression. Participants were instructed to read or review one lesson each week, and to complete the recommended homework assignment and additional reading and were prompted by approximately three automated emails each week, which provided instructions, reinforcement, and encouragement. Those in the therapist-guided groups received weekly, scripted, five to ten minute telephone or secure-email contact with a therapist, who checked progress, discussed strategies for overcoming problems in recovery from depression, and reinforced efforts. As there were no significant differences between the two courses or the self-guided or therapist-guided versions [22], data were analyzed from all treatment arms were analyzed together.

2.3. Measures

2.3.1. Patient Health Questionnaire, PHQ-9 [3]—The PHQ-9 is a widely used self-report measure of depressive symptom severity and is recommended by the fifth edition of the *Diagnostic and Statistical Manual of Mental Disorders* [26] for use to indicate symptom severity for major depressive disorder. The PHQ-9 was used in both trials and thus is the primary outcome for our analyses. In addition, the PHQ-9 is increasingly used in primary care, the de facto site for treatment of depression, to track depression [27]. Full remission

was defined as PHQ < 5, as this is an accepted criterion for full response [4, 28]. A raw symptom severity score was used, rather than a change score (sometimes used in research as an index of reliable change), as raw scores are both recommended as indices of response and more likely to be used in clinical practice [4, 28].

2.3.2. Hamilton Rating Scale for Depression, HAM-D [29]—The HAM-D is an interviewer administered 17-item rating scale of depressive symptom severity. The HAM-D was administered only in the CBT study, and was included to cross validate the PHQ-9. Remission was defined as HAM-D < 11, as this is an accepted criterion for response [28]. Clinical evaluators, blinded to treatment assignment and self-report outcomes, conducted the HAM-D interview. These evaluators had a minimum of a bachelor's degree and received training and supervision by a licensed psychologist. The HAM-D analyses were intended primarily for cross-validation of the PHQ-9 findings, as the HAM-D is not typically used in practice.

2.4. Statistical analyses

2.4.1. CBT study—Logistic regression models were fit for remission based on end-of-study (week 18) response to treatment (PHQ-9 <5 or HAM-D <11). We used PHQ-9 scores at treatment week 4, week 9, and week 14 to predict response. We first conducted analyses separately for each arm of the CBT study (telephone and face-to-face) but found similar results and thus combined patients from each arm into a single analysis.

2.4.2 I-CBT study—Initial logistic regression models were fit for remission based on end of treatment (week 9) response to treatment (PHQ-9 < 5) using PHQ-9 scores at treatment week 4. Upon further consideration that a PHQ-9 < 5 by week 9 may be an unrealistic outcome in a clinical setting and is inconsistent and much stricter than PHQ-9 < 5 at 18 weeks used in the CBT study, a secondary analysis considered improvement at week 9 as an outcome. To determine what threshold of PHQ-9 to use to denote improvement, an ROC analysis was conducted in the CBT study using week 9 PHQ-9 to predict remission at week 18. Youden's index [30], which considers the maximum sum of the sensitivity and specificity, found a PHQ-9 score of less than 9 would be the optimal threshold. We first conducted analyses separately for each condition of the I-CBT study (self-guided vs. therapist-guided and Depression Course vs. Wellbeing Course) and found similar results and thus combined patients from each condition in a single analysis.

2.4.3 Predicting Treatment Response—For both studies, ROC curves were produced in R version 3.0.1 [31]. We investigated optimal cut points that could be used as indicators that a patient would or would not reach remission by end of treatment. Sensitivity in this case was defined as the probability of predicting remission among those who did remit at the end of treatment (true positive rate); specificity as the probability of predicting non-remission for a person who did not remit (true negative rate). As predicting non-response might provide an opportunity to alter treatment in a therapeutic setting, we also present negative predictive values (NPV) or the probability of not remitting among those predicted to non-remit. In relation to clinical decision-making, high specificity would reflect how well we could predict not changing a treatment that would be effective, while high NPV would

reflect how those we identified as not likely to remit would actually fail to reach criteria for full remission. The choice of cut points may vary based on the context. For example, a stepped care paradigm, in which non-responders are switched to a more intensive treatment, may weight the importance of specificity over sensitivity to allow initial treatments an opportunity to work and preserve more costly treatments for those who are more definitively identified as at risk for non-response. Alternatively, monitoring for provider consideration of modest, cost neutral course corrections may weight sensitivity more heavily, relative to stepped care paradigms. We conducted these analyses assuming a stepped care paradigm in which treatment would be intensified for non-responders. As such, we preferenced high specificity and NPV over high sensitivity. We identified cut points summing sensitivity, specificity, and NPV and looking for maximum values that maintained a specificity of at least 90%. This is similar logic to that employed by Youden's index [30].

3. Results

We first conducted ROC analysis and calculated the AUC of these models in the CBT and I-CBT studies. Results from these analyses are presented in Table 1. AUC values of 0.7–.0.8 are considered acceptable, 0.8–0.9 excellent, and 0.9–1.0 outstanding [33]. All AUC values reported are in the acceptable range. In the CBT study, the overall rate of remission was 48% based on PHQ-9 <5. Table 2 displays the sensitivity, specificity, and NPV at each value of PHQ-9 scores for both the CBT and I-CBT studies. As these values represent empirical estimates, missing values are presented when no participants in the sample analyzed had a PHQ-9 score at that value. Based on our proposed balance of sensitivity, specificity, and NPV, values represented in grey would be unlikely to reach remission. The ROC analysis for week 4 showed that a cut point of PHQ-9 ≥17 provided a specificity (probability that PHQ-9 at week 4 was <17 among those who did remit) of 95% and a sensitivity of 32%. The NPV was 87%. With that cut point 19% of patients (n = 53) would be predicted to fail to reach remission by end of treatment. At week 9 (removing those who would already be predicted to fail to reach remission at treatment), we identified a cut point of PHQ-9 ≥ 13. This threshold provided 92% specificity and 30% sensitivity, with a NPV of 75%. This cut point resulted in 17% of patients (n = 40) predicted to fail to reach remission by end of treatment. At week 14, we identified a cut point of PHQ-9 ≥ 9 This cut point had a specificity of 91%, sensitivity of 40%, and NPV of 72%. Twenty-three percent of patients (n = 43) were predicted to fail to reach remission by end of treatment using that cut point.

In the I-CBT study, the overall rate of remission was 32%. The ROC analysis showed that a cut point of PHQ-9 ≥17 provided a specificity of 96% and sensitivity (probability that PHQ-9 at week 4 was ≥17 among those who failed to remit) of 16%. This resulted in the same criterion produced in the CBT study. The negative predictive value (NPV), or the probability of not remitting among those with PHQ-9 ≥17 at week 4 was 90%. If this threshold were used, 32% of patients would have been identified as at risk for non-remission.

In the CBT study, ratings of depression were also available on the HAM-D. We cross-validated the findings defining remission as HAM-D < 11. AUC values from the ROC curves were consistent with those found using PHQ < 5 as the outcome and were .71, .71,

and .74 for Weeks 4, 9, and 14 respectively. Additionally, the optimal cut points were consistent using HAM-D which is useful cross-validation for the findings because it is demonstrates PHQ-9 scores during treatment are predictive of remission on another measure of depression and one that uses clinical evaluator ratings rather than self-report.

4. Discussion

To our knowledge, this is the first study to empirically investigate optimal cut points for depressive symptoms on the PHQ-9 during the course of treatment to predict end of treatment response. We found that the specific values that optimized sensitivity and negative predictive value were consistent across trials of very different modalities suggesting consistency across very different modes of treatment. These values, however, were not consistent across time points. As length of time in treatment increases, lower levels of depressive symptom severity predict remission, consistent with the notion that response must be defined differently over time.

The consistency of cut points across trials is interesting both in terms of the confidence in the findings and their applicability to different kinds of clinical applications. This also speaks to the likelihood that these results might hold across additional treatment modalities, settings and populations, although more research is needed to investigate this, especially with modalities that might differ more than those used in the current study (e.g., medications).

Monitoring symptom change during treatment holds substantial value across various contexts of mental health treatment. At the individual level, therapists who track outcomes can make corrections during the course of therapy to improve the trajectory of an individual patient. These corrections might involve a tactical change such as altering the focus in a psychotherapeutic intervention or a strategic change such as switching to (or augmenting with) pharmacotherapy if psychotherapy was the current treatment. The goal of these alterations is to ensure the patient improves. At the system level, monitoring response to treatment and understanding the likelihood of eventual response can inform decision-making with implications for the eventual cost of the course of treatment. Cost-effectiveness is a major concern for systemwide implementation of therapeutic options and exploring indicators of treatment response can help inform decision-making and policies on service allocation. Identifying individuals who are on track for eventual remission or likely need a change in treatment strategy can help persevere more costly treatments for those who truly need them and reduce the costs for those who can benefit from a less-intensive therapeutic option. As such, it is critical to better have better tools to monitor patients, levels of clinical symptoms, and prognosis that can guide clinical decision-making.

Several limitations should be considered when interpreting our results. First, it is worth noting that although multiple measures, including diagnostic status, were used for eligibility in the CBT study, the I-CBT study was less stringent and selection of participants relied on specified values on the PHQ-9. This might introduce potential bias in terms of those recruited and therefore analyzed, although the PHQ-9 is used quite often in clinical practice and especially primary care [4,5]. Second, our protocol did not allow for flexibility in

aspects of the assessment that might exist in applied settings such as clinical practice. It could be that different cut points would emerge if different time periods were used (as we did find differences in the cut points for later time points). Third, length of treatment was fixed within each trial. Thus, although we found considerable similarity, it might not be the case if patients were treated until remission in the given treatment. Fourth, all of our treatments relied on CBT principles; it could be that different cut points would emerge with different therapeutic options such as pharmacotherapy. Furthermore, all individuals in this analysis included some form of treatment so it is impossible to determine what the trajectory would be in the absence of intervention. Nevertheless, these results are quite useful in the context of individuals receiving treatment. Fifth, we evaluated the use of cut points that are consistent with treatment guidelines that recommend continuing treatment until full remission is reached [34]. However, these cut points would not necessarily be applicable if other criteria, such as remission or percent change in symptoms are used. Lastly, we selected cut points based on criteria we believe would be reasonable for a stepped care paradigm, in which treatment is increased to a more intensive and costly modality. The relative importance of correctly identifying likely non-responders vs. avoiding misclassification non-responders of responders would vary depending on the reasons for monitoring, and would likely result in different cut points.

Our findings found optimal cut points using PHQ-9 scores in two trials of treatments for depression. Monitoring symptom change during treatment is critical for the optimization of treatment outcomes, as well as for defining procedures in stepped care and collaborative treatment programs. Further examination of the use of cut points and generalization of these findings to other treatment modalities such as medication or non-CBT based psychotherapy is warranted. Ultimately, it would be useful to determine if these cut points could be used meaningfully to create decision points in an intervention trial. As such, future work should aim to verify these cut points using a validation sample to determine if these cut points are pragmatically useful. This is a critical step to develop thresholds to be used clinically. Nevertheless, this study provides empirical guidance for the selection of cut points, when the PHQ-9 is used, that can predict response in the treatment of depression. We believe these findings can help inform treatment approaches that aim to improve outcomes and cost-effectiveness including sequenced treatment approaches, collaborative care, and stepped care.

References

1. Kessler RC, Berglund P, Demler O, Jin R, Koretz D, Merikangas KR, et al. The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R). JAMA. 2003; 289:3095–105. [PubMed: 12813115]
2. Basch E, Torda P, Adams K. Standards for patient reported outcome-based performance measures. JAMA. 2013; 310:139–40. [PubMed: 23839744]
3. Kroenke K, Spitzer RL, Williams JB, Lowe B. The Patient Health Questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. Gen Hosp Psychiatry. 2010; 32:345–59. [PubMed: 20633738]
4. MacArthur Foundation's Initiative. The Macarthur initiative on depression and primary care at Dartmouth and Duke: Depression management toolkit. Hanover, NH: Dartmouth; 2004.
5. Nease DE, Maloin JM. Depression screening: A practical strategy. J Fam Pract. 2003; 52:118–124. [PubMed: 12585989]

6. American Psychiatric Association. Practice guideline for the treatment of patients with major depressive disorder (revision). Am J Psychiatry. 2000; 157(4 Suppl):1–45.
7. Hollon SD, Thase ME, Markowitz JC. Treatment and prevention of depression. Psychol Sci Public Interest. 2002; 3:39–77. [PubMed: 26151569]
8. DeRubeis RJ, Hollon SD, Amsterdam JD, Shelton RC, Young PR, Salomon RM, et al. Cognitive therapy vs medications in the treatment of moderate to severe depression. Arch Gen Psychiatry. 2005; 62:409–16. [PubMed: 15809408]
9. Evans MD, Hollon SD, DeRubeis RJ, Piasecki JM, Grove WM, Garvey MJ, et al. Differential relapse following cognitive therapy and pharmacotherapy for depression. Arch Gen Psychiatry. 1992; 49:774–81. [PubMed: 1417429]
10. Lambert MJ, Harmon C, Slade K, Whipple JL, Hawkins EJ. Providing feedback to psychotherapists on their patients' progress: Clinical results and practice suggestions. J Clin Psychol. 2005; 61:165–74. [PubMed: 15609358]
11. Shimokawa K, Lambert MJ, Smart DW. Enhancing treatment outcome of patients at risk of treatment failure: Meta-analytic and mega-analytic review of a psychotherapy quality assurance system. J Consult Clin Psychol. 2010; 78:298–311. [PubMed: 20515206]
12. Marangell LB. Switching antidepressant for treatment-resistant major depression. J Clin Psychiatry. 2000; 62:12–7. [PubMed: 11575730]
13. Melfi CA, Chawla AJ, Croghan TW, Hanna MP, Kennedy S, Sredl K. The effects of adherence to antidepressant treatment guidelines on relapse and recurrence of depression. Arch Gen Psychiatry. 1998; 55:1128–32. [PubMed: 9862557]
14. Rush AJ, Trivedi MH, Wisniewski SR, Nierenberg AA, Stewart JW, Warden D, et al. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: A STAR*D report. Am J Psychiatry. 2006; 163:1905–17. [PubMed: 17074942]
15. Katon W, Unützer J, Wells K, Jones L. Collaborative depression care: History, evolution, and ways to enhance dissemination and sustainability. Gen Hosp Psychiatry. 2010; 32:454–64.
16. Katon W. Collaborative depression care models: From development to dissemination. Am J Prev Med. 2012; 42:550–52. [PubMed: 22516497]
17. National Institute for Health and Clinical Excellence. NICE Clinical Guidelines 90. London: NICE; 2009. Depression in adults. The treatment and management of depression in adults.
18. Bower P, Gilbody S. Stepped care in psychological therapies: access, effectiveness and efficiency – Narrative literature review. Br J Psychiatry. 2005; 186:11–7. [PubMed: 15630118]
19. van Straten A, Hill J, Richards DA, Cuijpers PA. Stepped care treatment delivery for depression: a systematic review and meta-analysis. Psychol Med. 2014:1–16.
20. Steidtmann D, Manber R, Blasey C, Markowitz JC, Klein DN, Rothbaum BO, et al. Detecting critical decision points in psychotherapy and psychotherapy + medication for chronic depression. J Consult Clin Psychol. 2013; 81:783–92. [PubMed: 23750462]
21. Mohr DC, Ho J, Duffecy J, Reifler D, Sokol L, Burns MN, et al. Effect of telephone-administered vs face-to-face-cognitive behavioral therapy on adherence to therapy and depression outcomes among primary care patients: a randomized trial. JAMA. 2012; 207:2278–85. [PubMed: 22706833]
22. Titov N, Dear BF, Johnston L, Lorian C, Zou J, Wootton B, Spence J, et al. Improving adherence and clinical outcomes in self-guided internet treatment for anxiety and depression: randomised controlled trial. PLoS ONE. 2013; 8:e62873. [PubMed: 23843932]
23. American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 4. Washington, DC: Author; 1994.
24. Mohr DC, Likosky W, Bertagnolli A, Goodkin DE, Van Der Wende J, Dwyer P, Dick LP. Telephone-administered cognitive-behavioral therapy for the treatment of depressive symptoms in multiple sclerosis. J Consult Clin Psychol. 2000; 68:356–61. [PubMed: 10780138]
25. Mohr DC, Hart SL, Julian L, Catledge C, Honos-Webb L, Vella L, et al. Telephone-administered psychotherapy for depression. Arch Gen Psychiatry. 2005; 62:1007–14. [PubMed: 16143732]
26. American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 5. Arlington, VA: American Psychiatric Publishing; 2013.

27. Regier DA, Narrow WE, Rae DS, Manderscheid RW, Locke BZ, Goodwin FK. The de facto US mental and addictive disorders service system: Epidemiologic Catchment Area prospective 1-year prevalence rates of disorders and services. Arch Gen Psychiatry. 1993; 50:85–94. [PubMed: 8427558]

28. Gelenberg, AJ.; Freeman, MP.; Markowitz, JC.; Rosenbaum, JF.; Thase, ME.; Trivedi, MH., et al. Practice guideline for the treatment of patients with major depressive disorder. 3. Washington, DC: American Psychiatric Association; 2010.

29. Hamilton M. Development of a rating scale for primary depressive illness. Br J Soc Clin Psychol. 1967; 6:278–96. [PubMed: 6080235]

30. Youden WJ. Index for rating diagnostic tests. Cancer. 1950; 3:32–5. [PubMed: 15405679]

31. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: 2013.

32. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receive operating characteristic curves: A nonparametric approach. Biometrics. 1988; 44:837–45. [PubMed: 3203132]

33. Hosmer, DW.; Lemeshow, S. Applied logistic regression. 2. Hoboken, NJ: John Wiley & Sons;

34. Karasu TB, Gelenberg A, Wang P, Merriam A, McIntyre JS, Charles SC, et al. Practice guideline for the treatment of patients with major depressive disorder (revision). Am J Psychiatry. 2000; 157(4 Suppl):1–45.

Table 1

ROC curves predicting response using PHQ-9 scores

Outcome	AUC (95%CI)*
CBT Study: Week 4	0.72 (0.66, 0.78)
CBT Study: Week 9	0.72 (0.66, 0.79)
CBT Study: Week 14	0.78 (0.72, 0.85)
I-CBT Study: Week 4	0.72 (0.64, 0.81)

* confidence interval calculated using DeLong's estimate of the variance of the AUC (DeLong et al., 1988).

Table 2

Empirical estimates of sensitivity, specificity, and NPV at each observed PHQ-9 score

	Week 4			Week 9			Week 14		
PHQ	**Sensitivity**	**Specificity**	**NPV**	**Sensitivity**	**Specificity**	**NPV**	**Sensitivity**	**Specificity**	**NPV**
CBT Study									
0	.99	.02	.53	.99	.09	.46	.97	.23	.43
1	.99	.05	.54	.98	.16	.48	.93	.37	.47
2	.98	.10	.54	.93	.26	.50	.88	.46	.50
3	.96	.18	.56	.89	.32	.51	.84	.59	.56
4	.93	.21	.56	.83	.45	.55	.78	.68	.60
5	.89	.32	.59	.76	.56	.58	.71	.77	.65
6	.82	.41	.61	.65	.67	.61	.56	.84	.67
7	.75	.51	.63	.57	.77	.66	.43	.89	.70
8	.69	.62	.66	.51	.82	.70	.40	.91	.72
9	.63	.65	.67	.43	.84	.68	.28	.92	.69
10	.60	.69	.68	.36	.86	.67	.21	.94	.68
11	.58	.78	.74	.33	.90	.72	.16	.96	.69
12	.52	.83	.77	.30	.92	.75	.11	.96	.61
13	.45	.84	.76	.22	.94	.73	.06	.96	.50
14	.40	.89	.81	.19	.94	.73	.04	.98	.60
15	.37	.92	.84	.11	.96	.69	.03	.98	.50
16	.32	.95	.87	.08	.98	.73	-	-	-
17	.25	.95	.86	.05	.98	.71	.01	.99	.50
18	.19	.98	.93	.05	1	1	0	.99	0
19	.16	.98	.92	.03	1	1	-	-	-
20	.12	.98	.89	-	-	-	-	-	-
21	.10	1	1	.01	1	1	-	-	-
22	.06	1	1	-	-	-	-	-	-
23	.01	1	1	-	-	-	-	-	-
24	.01	1	1	-	-	-	-	-	-
25	-	-	-	-	-	-	-	-	-

	Week 4			Week 9			Week 14		
PHQ	**Sensitivity**	**Specificity**	**NPV**	**Sensitivity**	**Specificity**	**NPV**	**Sensitivity**	**Specificity**	**NPV**
26	-	-	-	-	-	-	-	-	-
27	-	-	-	-	-	-	-	-	-
I-CBT Study									
0	-	-	-						
1	-	-	-						
2	1	.08	.70						
3	1	.17	.72						
4	.94	.25	.73						
5	.90	.31	.73						
6	.82	.44	.76						
7	.77	.52	.77						
8	.73	.60	.79						
9	.63	.65	.79						
10	.56	.75	.83						
11	.47	.83	.85						
12	.43	.86	.87						
13	.34	.94	.93						
14	.31	.94	.92						
15	.23	.94	.89						
16	.16	.96	.90						
17	.10	.96	.85						
18	.05	.96	.75						
19	.04	.96	.71						
20	.03	.98	.75						
21	.03	1	1						
22	-	-	-						
23	.02	1	1						
24	-	-	-						
25	-	-	-						
26	-	-	-						

	Week 4			Week 9			Week 14		
PHQ	Sensitivity	Specificity	NPV	Sensitivity	Specificity	NPV	Sensitivity	Specificity	NPV
27	-	-	-						