The Patient Health Questionnaire Somatic, Anxiety, and Depressive Symptom Scales: a systematic review

Kurt Kroenke, M.D.a,⁎, Robert L. Spitzer, M.D.b, Janet B.W. Williams, D.S.W.b, Bernd Löwe, M.D., Ph.D.c

aFrom Regenstrief Institute, Inc. and the Department of Medicine, Indiana University, Indianapolis, IN 46202, USA
bBiometrics Research Department, New York State Psychiatric Institute and Department of Psychiatry, Columbia University, New York, NY 10032, USA
cDepartment of Psychosomatic Medicine and Psychotherapy, University Medical Center Hamburg-Eppendorf and Schön Klinik, Hamburg-Eilbek, Germany

Received 4 October 2009; accepted 9 March 2010

Abstract

Background: Depression, anxiety and somatization are the most common mental disorders in primary care as well as medical specialty populations; each is present in at least 5–10% of patients and frequently comorbid with one another. An efficient means for measuring and monitoring all three conditions would be desirable.

Methods: Evidence regarding the psychometric and pragmatic characteristics of the Patient Health Questionnaire (PHQ)-9 depression, generalized anxiety disorder (GAD)-7 anxiety and PHQ-15 somatic symptom scales are synthesized from two sources: (1) four multisite cross-sectional studies (three conducted in primary care and one in obstetric-gynecology practices) comprising 9740 patients, and (2) key studies from the literature that have studied these scales.

Results: The PHQ-9 and its abbreviated eight-item (PHQ-8) and two-item (PHQ-2) versions have good sensitivity and specificity for detecting depressive disorders. Likewise, the GAD-7 and its abbreviated two-item (GAD-2) version have good operating characteristics for detecting generalized anxiety, panic, social anxiety and post-traumatic stress disorder. The optimal cutpoint is ≥10 on the parent scales (PHQ-9 and GAD-7) and ≥3 on the ultra-brief versions (PHQ-2 and GAD-2). The PHQ-15 is equal or superior to other brief measures for assessing somatic symptoms and screening for somatoform disorders. Cutpoints of 5, 10 and 15 represent mild, moderate and severe symptom levels on all three scales. Sensitivity to change is well-established for the PHQ-9 and emerging albeit not yet definitive for the GAD-7 and PHQ-15.

Conclusions: The PHQ-9, GAD-7 and PHQ-15 are brief well-validated measures for detecting and monitoring depression, anxiety and somatization.

1. Introduction

The Primary Care Evaluation of Mental Disorders (PRIME-MD) was an instrument developed and validated in the early 1990s to efficiently diagnose five of the most common types of mental disorders presenting in medical populations: depressive, anxiety, somatoform, alcohol and eating disorders [1]. Patients first completed a one-page 27-item screener and, for those disorders for which they screened positive, were asked additional questions by the clinician using a structured interview guide. The latter was modelled after lengthier structured psychiatric interviews which were useful in research but impractical in clinical practice [2–4]. The PRIME-MD proved to have good operating characteristics and seemed reasonably efficient: it took an average of 5.6 min of clinician time to administer the PRIME-MD to patients without a mental disorder diagnosis and 11.4 min to patients with a diagnosis.

⁎ Corresponding author. Tel.: +1 317 630 7447; fax: +1 317 630 6611. E-mail address: kkroenke@regenstrief.org (K. Kroenke).

0163-8343/$ – see front matter © 2010 Published by Elsevier Inc.
doi:10.1016/j.genhosppsych.2010.03.006
This modest investment of time was still a barrier to use given the competing demands in primary care where visits typically average 15 min or less and patients may have multiple acute and chronic medical disorders, preventive medicine needs and documentation requirements. Therefore, in two large studies enrolling 6000 patients (3000 from general internal medicine and family practice clinics and 3000 from obstetrics-gynecology clinics), a self-administered version of the PRIME-MD called the Patient Health Questionnaire (PHQ) was developed and validated [5,6]. In the past decade, the PHQ in general and the PHQ-9 depression scale in particular have gained increasing use in both research and practice.

Given the popularity of the PHQ-9 for assessing and monitoring depression severity, a new seven-item anxiety scale using a response set similar to the PHQ-9 was initially developed to diagnose generalized anxiety disorder (GAD) (hence its name, the GAD-7) and validated in 2740 primary care patients [7]. Though originally developed to diagnose generalized anxiety disorder, the GAD-7 also proved to have good sensitivity and specificity as a screener for panic, social anxiety and post-traumatic stress disorder [8]. Finally, the PHQ-15 was derived from the original PHQ studies and is increasingly used to assess somatic symptom severity and the potential presence of somatization and somatof orm disorders [9].

Each PHQ module can be used alone (e.g., the PHQ-9 if depression is the condition of interest), together with other modules or as part of the full PHQ. Although the PHQ was originally developed to detect five disorders, the depression, anxiety and somatoform modules (in that order) have turned out to be the most popular. Also, most primary care patients with depressive or anxiety disorders present with somatic complaints and co-occurrence of somatic, anxiety and depressive symptoms [the Somatic-Anxiety-Depressive (SAD) triad] is exceptionally common [10–14].

Thus, our article focuses on the PHQ-9 depression, GAD-7 anxiety, and PHQ-15 somatic symptom scales, drawing on data from both the original studies and other subsequent studies in the literature. We call this composite measure the Patient Health Questionnaire Somatic, Anxiety, and Depressive Symptom Scales (PHQ-SADS) (Appendix A). This review is particularly timely because of the large amount of clinical research published on the PHQ scales over the past decade, the frequent overlap of depressive, anxiety, and somatic symptoms, and the increasing emphasis as DSM-V develops to conduct dimensional as well as categorical assessments [15].

2. Literature search

We searched MEDLINE from 1999 through September 2009 using the following search terms: PRIME-MD, Patient Health Questionnaire, PHQ-9, PHQ-8, PHQ-2, GAD-7, PHQ-15. A total of 561 publications were identified, and the abstracts were reviewed (full bibliography is available upon request from the authors). Studies that gathered original data or that synthesized data from multiple studies as either a meta-analysis or a systematic review were retrieved and assessed for inclusion in our narrative review. Studies that enrolled small samples (typically <100 patients) or that focused on the full PHQ or on the alcohol or eating modules only were excluded. Since our review focuses in particular on the operating characteristics and utility of the PHQ depressive, anxiety and somatic symptom scales, we also cite only the best and representative studies when there are numerous papers on a secondary topic (use of the PHQ scales in patients with comorbid disorders or special populations, testing in another language or country, dissemination of the scales, etc.). Because of the breath of our topic, this article is a narrative review of the three scales informed by a comprehensive literature search rather than an exhaustive literature synthesis and meta-analysis.

3. Overview of key validation studies

Together, the four original validation studies represented nearly 10,000 patients. Participants in the three primary care studies had a mean age of 46–55 years old; 60–69% were women, 8–30% were African-American and 4–9% were Hispanic [1,5,7]. Participants in the obstetrics-gynecology study had a mean age of 31; 100% were women, 15% were African-American and 39% were Hispanic [6]. Table 1 summarizes key psychometric characteristics of the PHQ-9, GAD-7 and PHQ-15. In addition to the four original studies, there have been two meta-analyses of the PHQ-9 [16,17] and a large validation study of the PHQ-15 [18]. Selected findings from the pivotal studies in Table 1 along with other relevant articles are discussed in subsequent sections of this review.

4. Depressive symptoms

4.1. Diagnostic performance and psychometric characteristics

The PHQ-9 can be used either as a diagnostic algorithm to make a probable diagnosis of major depressive disorder (MDD) or as a continuous measure with scores ranging from 0 to 27 and cutpoints of 5, 10, 15 and 20 representing mild, moderate, moderately severe and severe levels of depressive symptoms. MDD should be considered in patients who endorse ≥5 of the 9 symptoms as present “more than half the days” (the 9th item counts if endorsed “several days”) and one of the first two symptoms (depressed mood or loss of interest) is endorsed. Most patients are able to complete all items (<5% items were missing in PHQ studies). If a self-administered PHQ-9 is subsequently reviewed in-person with the respondent, missing items may be probed, whereas a practical and diagnostically conservative approach to items left blank on
mailed or automated reports is to score blank items as 0. Another approach in some research studies has been to replace the missing value by the mean of the completed items if less than 20% of items are missing. This approach is less conservative but has a lower risk of missing persons — sensitivity to change of the PHQ-9 and HAM-D [32,33]. A five-point decline represents a clinically significant improvement [26]. A method for equating PHQ-9 severity scores to other depression measures has been reported [34]. Finally, a modified version of the PHQ-9 to assess a lifetime history of depression has also been developed [35].

### 4.2. Patients with medical comorbidity

The PHQ-9 has been used in clinical studies across a variety of medical conditions including neurological disorders [36,37], cardiovascular disease [38–42], HIV disease [43,44], diabetes [45,46], chronic kidney disease [47], cancer [48–50], rheumatological disorders [51,52], gastrointestinal disease [53], dermatological disorders [54], obstetrics-gynecology practices [6,55,56], ophthalmologic and otolaryngologic disorders [56,57], pain and other somatic symptoms [58–64], brain [65] and spinal cord injury [66] and a variety of chronic medical conditions [67,68]. In dissemination studies, the PHQ-9 has been favorably received by both primary care and mental health specialists [69,70].

### 4.3. Nursing home and cognitively impaired patients

The Minimum Data Set (MDS) is part of the US federally mandated process for clinical assessment of all residents in Centers for Medicare and Medicaid (CMS) certified nursing homes. To determine changes for the third revision (MDS 3.0), CMS sponsored a national validation study involving 3258 nursing home (NH) residents [71]. Most NH residents (86%) successfully completed the PHQ-9 interview, and NH staff were able to complete an observational version of the PHQ-9 for 424 (92%) of the 461 residents who did not complete the resident interview. There was excellent agreement between PHQ-9 results obtained independently.

---

**Table 1**

Psychometric characteristics of PHQ-9, GAD-7 and PHQ-15 derived from key studies

<table>
<thead>
<tr>
<th>Scale/Study</th>
<th>Sample size</th>
<th>Criterion Validity</th>
<th>Reliability</th>
<th>Area under Curve</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Sensitivity</td>
<td>Specificity</td>
<td>Internal (Cronbach α)</td>
</tr>
<tr>
<td>PHQ-9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kroenke [19]</td>
<td>6000 patients</td>
<td>.88</td>
<td>.88</td>
<td></td>
</tr>
<tr>
<td>Gilbody [16]</td>
<td>14 studiesb</td>
<td>.80</td>
<td>.92</td>
<td>–</td>
</tr>
<tr>
<td>Wittkampf [17]</td>
<td>4 studiesa</td>
<td>.77</td>
<td>.94</td>
<td>–</td>
</tr>
<tr>
<td>GAD-7</td>
<td>2740 patients</td>
<td>–</td>
<td>.92</td>
<td>–</td>
</tr>
<tr>
<td>Spitzer [7,8]</td>
<td></td>
<td>–</td>
<td>–</td>
<td>.80</td>
</tr>
<tr>
<td>PHQ-15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kroenke [9]</td>
<td>6000 patients</td>
<td>–</td>
<td>–</td>
<td>.80</td>
</tr>
<tr>
<td>Ravesteijn [18]</td>
<td>906 patientsc</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>.80</td>
<td>.80</td>
<td>.76</td>
</tr>
</tbody>
</table>

GAD = generalized anxiety disorder.

* For sensitivity and specificity, PHQ-9 studies used cutpoint ≥10 (except Gilbody who looked at both PHQ-9 diagnostic algorithm and cutpoint ≥10 and found no difference; GAD-7 study used cutpoint ≥10; and Ravesteijn used ≥3 severe symptoms (i.e., rated by patient as ‘bothered a lot’) on the PHQ-15.

b Gilbody and Wittkampf papers were meta-analyses of PHQ-9 diagnostic studies.

c For Ravesteijn study, line 1 are the psychometrics using PHQ-15 as a continuous score, and line 2 are psychometrics using a cutpoint of ≥3 severe symptoms.
by NH facility nurses and researchers (kappa=0.968). Moreover, compared to the 15-item Geriatric Depression Scale, the PHQ-9 required less time to complete, correlated more strongly with the criterion standard psychiatric assessment (.83 vs. .71), and showed more internal consistency across varying levels of cognitive ability. Most nurses rated the PHQ-9 as superior to depression assessment as currently performed in MDS 2.0 (88% of nurses), felt the PHQ-9 results could inform care plans (84%) and provide new insights into resident mood (86%). Most (77%) also reported that they felt that all residents who gave answers understood them (only 6% disagreed). Thus, the PHQ-9 is the depression measure in MDS 3.0. Another recent study found the PHQ-2 superior to four other measures (including the depression measure in MDS 3.0). The PHQ-9 is understood by NH facility nurses and researchers (kappa=0.968). The PHQ-A is one of the few depression measures specifically validated in adolescents [74]. The US Preventive Services Task Force recently recommended depression screening in adolescents (12–18 years of age) when systems are in place to ensure accurate diagnosis, psychotherapy (cognitive-behavioral or interpersonal) and follow-up. They concluded evidence is insufficient to warrant a recommendation to screen children (7–11 years of age). They advised that either the PHQ-A or the Beck Depression Inventory-Primary Care Version be used for screening. A different multidisciplinary expert consensus group produced the Guidelines for Adolescent Depression in Primary Care (GLAD-PC) report [75]. The GLAD-PC Toolkit suggests several different depression tools, one being a modified version of the PHQ-9 which preserves the four-item response set (and thus its continuous severity score) and added three additional items (two on suicidal ideation and one to screen for dysthymia). The traditional version of the PHQ-9 has recently proven sensitive to change in a small pilot study of adolescent depression [76]. Ideally, for studies that determine the operating characteristics of these different versions of the PHQ as well as alternative adolescent depression measures, the use of a criterion standard psychiatric interview would be desirable.

4.4. Adolescents

Johnson et al. developed the PHQ-A, a version of the PHQ modified for use in adolescent populations, and validated it in a sample of 403 primary care adolescent patients, 50 of whom had major depression as determined by an independent criterion standard psychiatric interview [23]. Compared to the PHQ-9, the PHQ-A depression module has more items (14 vs. 9), a yes–no response set (thus precluding a severity score) and much less cumulative validation data. However, the PHQ-A is one of the few depression measures specifically validated in adolescents [74]. The US Preventive Services Task Force recently recommended depression screening in adolescents (12–18 years of age) when systems are in place to ensure accurate diagnosis, psychotherapy (cognitive-behavioral or interpersonal) and follow-up. They concluded evidence is insufficient to warrant a recommendation to screen children (7–11 years of age). They advised that either the PHQ-A or the Beck Depression Inventory-Primary Care Version be used for screening. A different multidisciplinary expert consensus group produced the Guidelines for Adolescent Depression in Primary Care (GLAD-PC) report [75]. The GLAD-PC Toolkit suggests several different depression tools, one being a modified version of the PHQ-9 which preserves the four-item response set (and thus its continuous severity score) and added three additional items (two on suicidal ideation and one to screen for dysthymia). The traditional version of the PHQ-9 has recently proven sensitive to change in a small pilot study of adolescent depression [76]. Ideally, for studies that determine the operating characteristics of these different versions of the PHQ as well as alternative adolescent depression measures, the use of a criterion standard psychiatric interview would be desirable.

4.5. Pregnancy and postpartum period

Both pregnancy and the postpartum period are times of high-risk for the onset or recurrence of depression. The 10-item Edinburgh Postnatal Depression Scale (EPDS) is the most commonly-used depression scale for this population. Several studies have tested the PHQ-9 against a criterion standard psychiatric interview. In a longitudinal study of 506 women assessed at six time points over 9 months following delivery of their child, the PHQ-2 was found to be highly sensitive and the PHQ-9 was highly specific for identifying postpartum depression [77]. In head-to-head comparisons with the EPDS, the PHQ has been found to be more accurate (n=160 subjects) [78], comparable (n=415) [79], or slightly less accurate (n=123) [80]. The PHQ has recently been used in a depression screening initiative involving 1336 pregnant and postpartum women receiving obstetrical care at publicly funded health care clinics [81]. A recent evidence-based review of postnatal depression screening concluded that the optimal method to identify postnatal depression has not yet been identified [82].

4.6. Abbreviated versions: PHQ-2 and PHQ-8

The PHQ-2 consists of the first two items of the PHQ-9—depressed mood and loss of interest (anhedonia)—of which at least one is required to establish a diagnosis of MDD or any DSM-IV depressive disorder. These two items are comparable to many longer case-finding measures for depression screening whether asked with a yes–no response set as in the original PRIME-MD [83] or with the four-response set of the PHQ-9 [44,84–87]. PHQ-2 scores can range from 0 to 6, and a cutpoint ≥3 suggests clinically significant depression which should prompt either completion of the full PHQ-9 or a clinical interview to assess for MDD. The PHQ-2 also appears to be responsive when monitoring depression outcomes [86]. In a study of 1523 psychiatric outpatients evaluated in the Methods to Improve Diagnostic Assessment and Services project, the PHQ-2 items were found to be the most discriminatory for major depression from a databank of items [88].

The PHQ-8 omits the ninth item that asks about thoughts of death or self-harm. For large epidemiological studies or other research where depression is a secondary measure, some investigators have used the PHQ-8 because depression is not the main outcome, and it is just not feasible to have a back-up system for the very rare “suicidal” response to the ninth item. Even in clinical trials of depressed patients, when conducted in primary care or other medical populations rather than psychiatric patients, most patients who endorse the ninth item are agreeing with the first part of the item (passive thoughts of “being better off dead”) rather than the second part (active thoughts of “hurting yourself in some way”). The PHQ-8 has also been used in clinical or research settings where follow-up to positive responses to the ninth item of the PHQ-9 may be delayed (e.g., mailed or web-based screening).

The PHQ-8 has been shown to have comparable operating characteristics to the PHQ-9 in terms of diagnosing depressive disorders when using a DSM-IV based diagnostic algorithm [20,85]. Research indicates that deletion of the
ninth item has only a minor effect on scoring because it is, by far, the least frequently endorsed item on the PHQ-9 [20,28,89,90]. In 1004 primary care patients, the correlation between PHQ-9 scores and PHQ-8 scores was very high ($r=.998$); only three patients with a PHQ-9 score of 10 or higher had a PHQ-8 score of less than 10 [85]. In our two PHQ validation studies totaling 6000 patients, we have confirmed this (data not previously published) by finding very high correlations between the PHQ-9 and PHQ-8 ($r=.997$ in both samples) and a similar area under the curve (0.95) by receiver operating characteristic (ROC) analysis for the two measures in diagnosing major depression. Thus, identical cutoffs can be used on the PHQ-8 and PHQ-9 for both clinical and research purposes [91].

4.7. Alternative response sets

In some settings, the response set has been converted from the standard verbal options to number of days in the past 1 or 2 weeks. This has facilitated use of the PHQ in certain survey settings [91] as well as for telephone-based automated depression monitoring by interactive voice recording [49]. Data on MDD prevalence rates and construct validity suggest that the number of days response set may be an acceptable option [91]. Since the standard response set has much stronger validation data, however, direct comparisons of the two response sets in the same population would be advisable before widespread adoption of the number of days alternative.

When respondents are asked to report the number of days they have been bothered by each PHQ-9 symptom in the past 2 weeks, the PHQ-9 scores for 0–1, 2–6, 7–11 and 12–14 days are 0, 1, 2 and 3, respectively [91]. The rationale is that 0–1 days is clearly less than “several days” and probably not clinically significant; 7–11 days is “more than half the days” and the training manual for the Structured Clinical Interview for DSM-IV Axis I Disorders instructs the interviewer to consider more than 11 days as the cutoff if respondents insist on explicit guidance for what is meant by “nearly every day in the past 2 weeks” [92]. When a 1-week time frame is used, the PHQ-9 scores for 0–1 days, 2–3 days, 4–5 days and 6–7 days are 0, 1, 2 and 3, respectively [49].

4.8. Dissemination

The PHQ-9 has had a considerable amount of uptake both clinically and by researchers in less than a decade. Examples include its:

- Selection as the clinical tool used in the intervention arm of numerous randomized effectiveness trials in medical populations for monitoring and adjusting depression therapy [32,49,50,93–100]
- Use in its full or abbreviated forms in large federally-sponsored surveys and programs including the Behavioral Risk Factor Surveillance Survey [101]; the National Health and Nutrition Examination Survey [102]; the Medical Expenditure Panel Survey [103]; the National Epidemiologic Survey on Alcohol and Related Conditions [87]; the Medicare Health Support program [104], the Millennium Cohort Study [105] and population-based studies in Germany [106], Australia [107] and Canada [108].
- Adoption as a standard measure for depression screening by the Veterans Administration [109], Department of Defense [110], Kaiser [111] and several integrated health care systems [112], managed behavioural care organizations [113] and public health departments [114].
- Popularity as the most commonly used depression measure in the United Kingdom’s National Health Service which requires use of a measure as part of depression treatment in primary care [115].
- Incorporation into toolkits or programs to improve depression care by the MacArthur Foundation [116] and the Robert Wood Johnson Foundation [117] as well as professional organizations like the American Heart Association [118].
- Consideration by the American Psychiatric Association as the dimensional depression measure in its DSM-V classification manual [119].
- Acceptability, utility and sustainability in psychiatry and primary care practice networks [69,70,120].

5. Anxiety symptoms

5.1. Original PHQ anxiety module

The original PHQ anxiety module focused on two diagnoses: panic disorder and other anxiety disorder. The 15-item panic module yielded a probable diagnosis of panic disorder for individuals who answered “yes” to the first four questions and endorsed ≥4 of 11 somatic symptoms during an anxiety attack. In addition to validation in the two original PHQ studies, further research has strengthened the evidence for the panic disorder section [89,121–125]. The other anxiety disorder section primarily included criteria for GAD but its yes-no response format did not permit calculation of a severity score that proved so useful with the PHQ-9. Thus, a study in more than 2700 primary care patients was conducted to develop and validate a 7-item anxiety measure (GAD-7) with a similar response set to the PHQ-9 in order to establish probable diagnoses of GAD as well as grade its severity [7].

5.2. Diagnostic performance and psychometric characteristics of GAD-7 for GAD

GAD-7 scores can range from 0 to 27, with 5, 10 and 15 representing mild, moderate and severe levels of anxiety symptoms [7,8]. Psychometric characteristics of the GAD-7 are summarized in Table 1. At a cutoff of ≥10, both sensitivity and specificity exceeded .80, and sensitivity was
nearly maximized. Scale characteristics were not influenced by age, sex or race/ethnicity. However, the proportion of primary care patients who score at this level was quite high (23%). Thus, a cutoff of ≥15 maximizes specificity and approximates a prevalence (9%) more in line with current epidemiologic estimates of GAD prevalence in primary care. However, sensitivity at this high cutpoint is low (48%).

Although the GAD-7 inquires about symptoms in the past 2 weeks, the DSM-IV diagnostic criteria for GAD specify at least a 6-month duration of symptoms. Nonetheless, the operating characteristics of the scale were good because most patients with high symptom scores had in fact chronic symptoms. Of the 433 patients with GAD-7 scores ≥10, 96% had symptoms ≥1 month and 67% had symptoms ≥6 months. Notably, the National Comorbidity Survey has shown that cases with episodes of GAD for 1–5 months do not differ greatly from those with episodes of ≥6 months in onset, persistence, impairment, co-morbidity, parental GAD or sociodemographic correlates [126]. Kessler et al. conclude that there is little basis for excluding people from a diagnosis of GAD based simply on duration of symptoms.

There was a strong association between increasing GAD-7 severity scores and worsening function on all six Medical Outcomes Study Short-Form 20-item General Health Survey (SF-20) health-related quality of life scales [7]. Convergent validity of the GAD-7 was good, as demonstrated by its correlations with two anxiety scales: the Beck Anxiety Inventory (r=.72) and the anxiety subscale of the SCL-90 (r=.74). Factor analysis that included the GAD-7 anxiety items and the PHQ-8 depression items confirmed two distinct dimensions, with all depression items having the highest factor loadings on one factor (0.58–0.75) and all anxiety items having the highest factor loadings on the second factor (0.69–0.81). Although sensitivity to change of the GAD-7 is suggested by the results of a large randomized depression trial [100] where anxiety was a secondary outcome, responsiveness will be better established in treatment trials or other types of longitudinal studies where anxiety is the primary outcome.

The validity of the GAD-7 for population-based epidemiologic studies is supported by results from a nationally representative face-to-face household survey of 5030 subjects conducted in Germany [127]. Factor analysis substantiated the 1-dimensional structure of the GAD-7 and its factorial invariance for gender and age. Internal consistency was identical across all subgroups (0.89). Approximately 5% of subjects had GAD-7 scores of 10 or greater and 1% had GAD-7 scores of 15 or greater.

5.3. Operating characteristics of GAD-7 for other anxiety disorders

The four most common anxiety disorders (excluding simple phobias which seldom present clinically) are generalized anxiety disorder, panic disorder, social anxiety disorder and posttraumatic stress disorder (PTSD) [128]. In the GAD-7 validation study, PTSD was present in 8.6% of the patients, GAD in 7.6%, panic disorder in 6.8%, social anxiety disorder in 6.2% and any anxiety disorder in 19.5%, all of which are in the prevalence range reported in previous primary care studies.

Although originally developed for GAD, the GAD-7 also proved to have good sensitivity and specificity as a screener for panic, social anxiety and post-traumatic stress disorder [8]. Not surprisingly, the area under the curve by ROC analysis is greatest (.91) for GAD, but also quite good for panic disorder (.85), social anxiety disorder (.83) and PTSD (.83). Each of the four anxiety disorders was strongly and similarly associated with impaired functioning on all six SF-20 scales and with self-reported disability days. In fact, the number of anxiety disorders rather than the specific type of disorder was the factor most strongly associated with impairment.

5.4. Abbreviated versions: GAD-2

The GAD-2 consists of the first two items of the GAD-7 which in turn correspond to the two core diagnostic criteria for GAD. Scores on the GAD-2 range from 0 to 6, and a cutpoint ≥3 denotes a screening cutpoint for clinically significant anxiety which should prompt completion of the full GAD-7 and a clinical interview to determine the type of anxiety disorder and whether treatment and/or referral is warranted [8]. Notably, the area under the curve by ROC analysis for the GAD-2 was similar to that for the GAD-7 for three of the four anxiety disorders and only slightly lower (.80 vs. .83) for PTSD.

5.5. Dissemination

Compared to the PHQ-9, uptake of the GAD-7 is in a more nascent phase given its later publication (2006 vs. 1999). Some experts are recommending its routine use [129], and new research studies using the GAD-7 are starting to emerge [49,100,130,131]. Unlike the PHQ-9 which can serve as both a diagnostic and severity measure, the GAD-7 is principally a measure of anxiety severity; the likelihood of an anxiety disorder increases with higher GAD-7 scores, but a clinical interview is required to confirm the presence and type of disorder.

6. Somatic symptoms

6.1. Three limitations in the classification of somatoform disorders

Three findings from recent research regarding the classification of somatoform disorders is relevant to the PHQ-15 [132,133]. First, fewer chronic symptoms are probably needed than has traditionally been required for the diagnosis of somatization disorder which captures less than 10–20% of the patients with chronic and disabling somatization in primary care. Second, focusing primarily on
current rather than lifetime symptom counts may be desirable due to the greater reliability and comparable validity of current counts. Third, the requirement that symptoms be “medically unexplained” is problematic since confidently attributing the etiology of pain, fatigue, dizziness, gastrointestinal complaints and numerous other common somatic symptoms to specific medical or psychological conditions that frequently coexist in the same patient is often difficult. Even disease-specific physical symptoms in patients with medical disorders may be explained as much by comorbid depression or anxiety as by the severity of the medical disorder [134]. By allowing lower symptom thresholds, focusing on current symptoms and not requiring adjudication of symptom etiology, the PHQ-15 addresses these three issues [135].

6.2. Diagnostic performance and psychometric characteristics of the PHQ-15

The PHQ-15 includes 15 symptoms that account for more than 90% of symptoms seen in primary care (exclusive of upper respiratory symptoms such as cough, nasal symptoms, sore throat, ear ache, etc.). [9]. The PHQ-15 asks patients to rate how much they have been bothered by each symptom during the past month on a 0 (“not at all”) to 2 (“bothered a lot”) scale. Thus, the total score ranges from 0 to 30, with cutpoints of 5, 10 and 15 representing thresholds for mild, moderate and severe somatic symptom severity, respectively. In the original study, the majority (88%) of patients who endorsed ≥3 medically unexplained symptoms at the level of “bothered a lot” and who had at least a several year history of poorly explained symptoms had a somatoform diagnosis [136,137]. In a recent study of the PHQ-15 in 906 primary care patients, a cutpoint of ≥3 severe (i.e., “bothered a lot”) symptoms during the past 4 weeks had a sensitivity of 78% and a specificity of 71% for a DSM-IV somatoform diagnosis [18]. These somewhat lower operating characteristics may be because the PHQ-15, unlike DSM-IV, asks only about current (rather than lifetime) symptoms and does not adjudicate whether or not a symptom is medically unexplained. Indeed, both DSM-IV criteria (lifetime recall and symptom attribution) have suboptimal reliability and validity in diagnosing somatoform disorders [132,133].

In the original PHQ studies of 6000 unselected primary care patients, higher PHQ-15 scores were strongly associated with worsening function on all six SF-20 scales as well as increased disability days and health care utilization [9]. In a smaller study of 172 primary care patients with at least moderate somatization enrolled in a clinical trial, Cronbach’s alpha was 0.79 and there was a moderately strong correlation (r=0.52) between PHQ-15 scores and medically unexplained symptom counts elicited by an independent structured psychiatric interview [138]. There is some support for the PHQ-15’s sensitivity to change in clinical trials or longitudinal studies, both as a primary [139–141] and a secondary [142,143] outcome measure.

There are several factors arguing for the PHQ-15 as an excellent measure of somatic symptom burden and potential somatization [135,144]. First, approximately 10% of primary care and obstetric-gynecology outpatients have a score of 15 or greater, a prevalence consistent with other studies of clinically significant somatization. Second, increasing scores on the PHQ-15 are strongly associated with functional impairment, disability, and health care use. Third, items on the PHQ-15 overlap better with other validated somatization screeners than any other two screeners do with one another. Fourth, it is an excellent measure for identifying high-utilizing somatizing patients in health care systems [145,146]. For example, patients with somatization as identified by the PHQ-15 had approximately twice the outpatient and inpatient medical care utilization and twice the annual medical care costs of nonsomatizing patients. Adjusting the findings for the presence of psychiatric and medical comorbidity had relatively little effect on this association. Fifth, total self-reported PHQ somatic symptom counts have been shown to be highly associated with clinician-rated somatoform disorder symptom counts [138,147].

6.3. Dissemination

Uptake of the PHQ-15 to date has largely been in the area of research, in part due to lower clinician interest in chronic symptoms and somatoform disorders as well as skepticism about effective treatments [148,149]. However, recent reviews have identified effective pharmacological and behavioral interventions [150–152]. Besides the studies already cited, examples of other major studies where the PHQ-15 has been a key measure include:

- An international study of mental disorders in primary care conducted in 15 countries that used the earlier yes-no PRIME-MD version of the PHQ-15 [11];
- A prospective study of physical symptoms in predicting hospitalization and mortality in 3498 elderly patients [153];
- A study of somatization in 10,507 consecutive patients attending 340 Australian general practices [154];
- A study of somatization in a representative sample (n=2510) of the German general population [155];
- A 20-year cohort study of more than 100,000 military personnel [156];
- A 10-year follow-up study of 30,000 veterans from the Gulf War era [157];
- A study of traumatic brain injury in 2525 US soldiers returning from Iraq [158];
- A primary care study of the impact of somatization on disability [159];

The PHQ-15 has also been used as a secondary measure in a variety of smaller studies examining mental disorders or physical symptom syndromes.
7. SAD triad

7.1. Overlap and additive effects

The comorbidity of somatic, anxiety and depressive symptoms (the “SAD” triad) is well-established [12,13]. Three recent large epidemiologic studies in primary care confirm that “pure” forms are much less common than overlapping syndromes; most patients reporting high levels of one symptom type also report high levels of one or both of the other types of symptoms [14,154,160]. Also, somatic, anxiety and depressive symptoms have independent, additive and differential effects on multiple domains of health-related quality of life, functional status, disability and health care use [9,14,161,162].

7.2. Incremental impairment

Table 2 summarizes data on health-related quality of life (HRQL) and symptom severity from the original studies [7,9,19]. When moving across the ordinal categories of minimal to mild to moderate to severe scores, there is a large and progressive decrement across multiple domains of HRQL for all three types of symptoms. For simplicity, only mean scores are displayed (standard deviations are provided in the original papers); however, most pair-wise comparisons of mean SF-20 scores between adjacent symptom severity levels within each scale are highly significant (P<.001). Figs. 1 and 2 show a similar relationship between increasing severity level and self-reported symptom-related difficulty and disability days for all three scales.

7.3. Similar cutpoints

Scores of 0–4, 5–9, 10–14, and ≥15 represent minimal, mild, moderate, and severe levels of symptom burden, respectively, on the PHQ-9, GAD-7, and PHQ-15. This results in a simple 5-10-15 cutpoint rule-of-thumb for all three scales. Also, a score ≥10 is the most commonly-recommended cutpoint for “clinically significant” symptoms on all three scales. Finally, a score ≥3 is the suggested cutpoint for the PHQ-2 and GAD-2.

However, inflexible adherence to a single cutpoint should be discouraged. For one thing, the maximum scores are not identical: 21, 27, and 30 for the GAD-7, PHQ-9, and PHQ-15, respectively. More importantly, one might choose a different cutpoint depending upon the population being assessed [23,163](community vs. primary care vs. mental health setting) and the purpose of the

---

Table 2

<table>
<thead>
<tr>
<th>Level of Symptom Severity</th>
<th>PHQ score (% of Dep/Anx/Som)</th>
<th>Mean SF-20 score* by depression (Dep), anxiety (Anx), and somatic (Som) symptom severity level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dep</td>
<td>Anx</td>
</tr>
<tr>
<td>Minimal 0–4 (62/56/35)</td>
<td>81</td>
<td>82</td>
</tr>
<tr>
<td>Low 5–9 (23/24/35)</td>
<td>66</td>
<td>65</td>
</tr>
<tr>
<td>Medium 10–14 (8/12/20)</td>
<td>52</td>
<td>54</td>
</tr>
<tr>
<td>High ≥15 (7/8/9)</td>
<td>39</td>
<td>41</td>
</tr>
</tbody>
</table>

* Each SF-20 scale is scored from 0 (worst) to 100 (best) health-related quality of life on that domain. SF-20 scores are adjusted for age, sex, education for all scales, and also number of physical disorders for Dep and Som.

---

Fig. 1. Proportion of patients with minimal, mild, moderate, and severe levels of symptoms who respond with either “very difficult” or “extremely difficult” to the last question of the PHQ: “How difficult have these problems made it for you to do your work, take care of things at home, or get along with other people?” Somatic symptoms are measured by the PHQ-15, anxiety symptoms by the GAD-7 and depressive symptoms by the PHQ-9. On all three scales, minimal symptoms are represented by a score of 0 to 4, mild symptoms by a score of 5–9, moderate symptoms by a score of 10–15 and severe symptoms by a score ≥15. Data is derived from the three original validation papers [7,9,19].
assessments (routine screening vs. evaluating suspected cases). Also, modest nuances of the ≥10 cutpoint has been suggested by some investigators, such as ≥12 for the PHQ-9 [22,115,164] and ≥8 for the GAD-7 [8]. Indeed, these latter cutpoints for the PHQ-9 and GAD-7 are the ones being recommended as part of the United Kingdom’s primary care Quality and Outcomes Framework (QOF) [115] and Improving Access to Psychological Therapies (IAPT) program [165], notable in light of the large amount of routine data these two initiatives will generate regarding the PHQ-9 and GAD-7.

7.4. Patient-rated difficulty item

The final question on the PHQ asks the patients to report “how difficult have these problems made it for you to do your work, take care of things at home, or get along with other people?” This single patient-rated difficulty item is not used in calculating any PHQ score or diagnosis but rather represents the patient’s global impression of symptom-related impairment. It may be useful in decisions regarding initiation of or adjustments to treatment since it is strongly associated with both psychiatric symptom severity as well as multiple measures of impairment and HRQL. In the two large validation studies [5,6], the percentage of subjects with a PHQ diagnosis varied significantly (P<.001) based upon response to this single item, ranging from 16% of the subjects who responded “not difficult at all,” to 35% who responded “somewhat difficult,” to 73% who responded “very difficult,” to 88% who responded “extremely difficult.” This question was also associated with functional impairment: the mean correlation of this item with each of the 6 SF-20 scales was .36, with the highest being 0.51 for mental health.

7.5. Ultra-brief scales

Scales as short as 2–4 items have good operating characteristics for depression screening and appear to perform as well as longer case-finding measures [166]. Validated individually as abbreviated screeners for depression and anxiety, the PHQ-2 and GAD-2 have also been validated when combined as an ultra-brief screener — the PHQ-4 — for depression and anxiety in large clinical (n=2149) [167] and general population (n=5030) samples [168]. The construct and factorial validity of the two-item depression and two-item anxiety subscales was confirmed, as was the recommended cutpoint of ≥3 on each subscale. Depression and anxiety scores were shown to have independent effects on impairment yet also had synergistic value in identifying patients with single or dual disorders: of the 2013 primary care patients who completed the PHQ-4 12.1% had anxiety only (GAD-2 score ≥3), 4.4% had depression only (PHQ-2 score ≥3), and 8.5% had both anxiety and depression [167].

7.6. PHQ-SADS

The PHQ-SADS (Appendix A) includes the PHQ-15, GAD-7 and PHQ-9 scales, plus the first five items of the panic module. A positive response to the first panic question alone has a sensitivity and specificity of 93% and 78%, respectively, while each additional “yes” response to the remaining four items increases specificity with only a minimal decline in sensitivity [121]. Because the 11 somatic symptoms in the original PHQ only marginally improve the operating characteristics of the panic module, they are omitted from the PHQ-SADS for the sake of brevity. The final item on the PHQ-SADS is the respondent’s global rating of symptom-related difficulty. As noted, scores of 5, 10 and 15 indicates increasing levels of severity on all three scales, and elevated scores on two or more scales suggests comorbidity. Responses to the single-item difficulty question can further guide treatment decisions.

7.7 Efficiency and acceptability

In the two PHQ validation studies totaling 6000 patients, the median time for physician review of the full PHQ after completion by a patient was less than 2 min [5,6]. Since the full PHQ assesses two additional disorders (substance and eating), the PHQ-SADS should take even less time. In these same two studies, 87–89% of physicians found the PHQ very or somewhat useful in management or treatment planning, 88–93% of patients felt very or somewhat comfortable answering the PHQ questions and 89–93% of patients believed the questions were very or somewhat helpful in getting their doctor to better understand or treat the problems that they were having.
8. Conclusion

While we have summarized the salient PHQ data, a comprehensive comparison of the PHQ with all alternative scales is beyond the scope of this review. Evidence-based literature syntheses of depression measures have recently been published [21,166]. Another limitation is the uncertain efficacy of screening for mental disorders on outcomes. For example, it appears that depression screening alone may not be sufficient to improve patient outcomes [169]. On the other hand, there is substantial evidence that when screening is part of a multi-component intervention, depression outcomes are improved [170,171]. That is why the US Services Preventive Services Task Force recommends routine screening for depression only if there are systems in place to deliver adequate treatment and follow-up [172]. A similar degree of evidence has not yet accumulated to justify routine screening for anxiety or somatoform disorders.

Nonetheless, all three conditions are sufficiently prevalent, disabling and costly in both primary care and specialty clinics to trigger case detection at a relatively low index of suspicion. Brief, self-administered scales are an efficient method for stratifying patients into screen-negative and screen-positive groups, thus allowing busy clinicians to prioritize their limited interview time for the smaller group of patients with high scores. Scales are also efficient tools for monitoring response to treatment; measurement-based care has been shown to enhance depression outcomes [173] and is likely to be beneficial for anxiety and somatization as well. Although the uptake of standardized scales into clinical practice has been slow [174,175], the potential benefits are reviewed elsewhere [175,176], including avoiding over- or undertreatment, monitoring response, adjusting treatment, assessing quality of care and standardizing communication among clinicians regarding disease severity.

Finally, scales are useful for identifying comorbidity, e.g., the somatizing depressed patient, or the patient with mixed anxiety-depression. This can be especially important in patients not responding to standard treatments, e.g., the depressed patient with persistent anxiety or pain who may need additional therapies. The PHQ depressive, anxiety and somatic symptom scales that have been developed, studied and clinically applied over the past decade constitute valid and efficient instruments for detecting, differentiating and monitoring the SAD triad. The PHQ-SADS (Appendix A) combines all three scales to provide a continuous severity measure of each of these three common and overlapping symptom domains. Such dimensional assessment is being advocated in the development of DSM-V to complement categorical diagnoses [15]. Finally, translations of the PHQ scales are currently available in more than 60 languages which facilitate their use in studying mental disorders and improving clinical outcomes on a global level.

Appendix A. The Patient Health Questionnaire Somatic, Anxiety, and Depressive Symptom Scales (PHQ-SADS)

Patient Health Questionnaire (PHQ-SADS)

This questionnaire is an important part of providing you with the best health care possible. Your answers will help in understanding problems that you may have. Please answer every question to the best of your ability.

A. During the last 4 weeks, how much have you been bothered by any of the following problems?

<table>
<thead>
<tr>
<th>Problem</th>
<th>Not bothered (0)</th>
<th>Bothered a little (1)</th>
<th>Bothered a lot (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stomach pain</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
</tr>
<tr>
<td>Back pain</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
</tr>
<tr>
<td>Pain in your arms, legs, or joints (knees, hips, etc.)</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
</tr>
<tr>
<td>Feeling tired or having little energy</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
</tr>
<tr>
<td>Trouble falling or staying asleep, or sleeping too much</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
</tr>
<tr>
<td>Menstrual cramps or other problems with your periods</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
</tr>
<tr>
<td>Pain or problems during sexual intercourse</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
</tr>
<tr>
<td>Headaches</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
</tr>
<tr>
<td>Chest pain</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
</tr>
<tr>
<td>Dizziness</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
</tr>
<tr>
<td>Fainting spells</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
</tr>
<tr>
<td>Feeling your heart pound or race</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
</tr>
<tr>
<td>Shortness of breath</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
</tr>
<tr>
<td>Constipation, loose bowels, or diarrhea</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
</tr>
<tr>
<td>Nausea, gas, or indigestion</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
</tr>
</tbody>
</table>

PHQ-15 Score □□□□□□□□□□□□□□□□ = □□□□□□□□□□□□□□□□ + □□□□□□□□□□□□□□□□

B. Over the last 2 weeks, how often have you been bothered by any of the following problems?

<table>
<thead>
<tr>
<th>Problem</th>
<th>Not at all (0)</th>
<th>Several days (1)</th>
<th>More than half the days (2)</th>
<th>Nearly every day (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feeling nervous anxiety or on edge</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
</tr>
<tr>
<td>Not being able to stop or control worrying</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
</tr>
<tr>
<td>Worrying too much about different things</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
</tr>
<tr>
<td>Trouble relaxing</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
</tr>
<tr>
<td>Being so restless that it is hard to sit still</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
</tr>
<tr>
<td>Becoming easily annoyed or irritable</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
</tr>
<tr>
<td>Feeling afraid as if something awful might happen</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
<td>□□□□□□□</td>
</tr>
</tbody>
</table>

GAD-7 Score □□□□□□□□□□□□□□□□ = □□□□□□□□□□□□□□□□ + □□□□□□□□□□□□□□□□ + □□□□□□□□□□□□□□□□

C. Questions about anxiety attacks.

a. In the last 4 weeks, have you had an anxiety attack — suddenly feeling fear or panic? YES NO
b. If you checked “NO”, go to question E. □□□□□□□ □□□□□□□
   a. Has this ever happened before? □□□□□□□ □□□□□□□
c. Do some of these attacks come suddenly out of the blue — that is, in situations where you don’t expect to be nervous or uncomfortable? □□□□□□□ □□□□□□□
D. Over the last 2 weeks, how often have you been bothered by any of the following problems?

- 1. Little interest or pleasure in doing things
- 2. Feeling down, depressed, or hopeless
- 3. Trouble falling or staying asleep, or sleeping too much
- 4. Feeling tired or having little energy
- 5. Poor appetite or overeating
- 6. Feeling bad about yourself — or that you are a failure or that you are a disgrace to your family down
- 7. Trouble concentrating on things, such as reading the newspaper or watching television
- 8. Moving or speaking so slowly that other people could have noticed? Or the opposite — being so fidgety or restless that you have been moving around a lot more than usual
- 9. Thoughts that you would be better off dead of or that you wish you were never born
- 10. Being so fidgety

PHQ-9 Score □□□□ + □□□□ + □□□□ = □□□□

E. If you checked off any problems on this questionnaire, how difficult have these problems made it for you to do your work, take care of things at home, or get along with other people?

□ Not difficult at all □ Somewhat difficult □ Very difficult □ Extremely difficult

Developed by Robert L. Spitzer, Janet B.W. Williams, Kurt Kroenke, and colleagues, with an educational grant from Pfizer. The names PRIME-MD and PRIME-MD TODAY are trademarks of Pfizer.

References


Patten SB, Schopflocher D. Longitudinal epidemiology of major depression as assessed by the Brief Patient Health Questionnaire (PHQ-9). Comprehensive Psychiatry 2009;50:26–33.

Desai MM, Rosenheck RA, Craig TJ. Case-finding for depression among medical outpatients in the Veterans Health Administration. Med Care 2006;44:75–81.


