

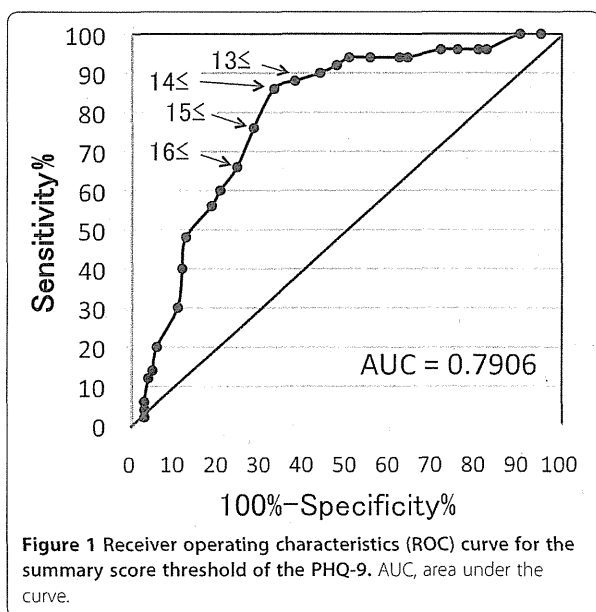
Table 2 Sensitivity, specificity, positive and negative predictive values, and positive and negative likelihood ratios, and overall accuracy for the diagnostic algorithm and the summary score (≥ 10) of the PHQ-9

	Diagnosis of current major depressive episode (MDE)		Diagnosis of MDE with major depressive disorder	
	Diagnostic algorithm of PHQ-9	Summary score (≥ 10) of the PHQ-9	Diagnostic algorithm of PHQ-9	Summary score (≥ 10) of the PHQ-9
True positive, <i>n</i> (%)	39 (25)	47 (31)	32 (21)	38 (25)
False positive, <i>n</i> (%)	34 (22)	52 (34)	41 (27)	61 (40)
False negative, <i>n</i> (%)	11 (7)	3 (2)	9 (6)	3 (2)
True negative, <i>n</i> (%)	69 (45)	51 (33)	71 (46)	51 (33)
Sensitivity (95%CI)	0.78 (0.64–0.88)	0.94 (0.84–0.99)	0.78 (0.62–0.89)	0.93 (0.80–0.98)
Specificity (95%CI)	0.67 (0.57–0.76)	0.50 (0.40–0.60)	0.63 (0.54–0.72)	0.46 (0.36–0.55)
Positive predictive value (95%CI)	0.53 (0.41–0.65)	0.47 (0.37–0.58)	0.44 (0.32–0.56)	0.38 (0.29–0.49)
Negative predictive value (95%CI)	0.86 (0.77–0.93)	0.94 (0.85–0.99)	0.89 (0.80–0.95)	0.94 (0.85–0.99)
Positive likelihood ratio (95%CI)	2.36 (1.76–2.99)	1.86 (1.56–2.02)	2.13 (1.59–2.63)	1.70 (1.41–1.85)
Negative likelihood ratio (95%CI)	0.33 (0.19–0.53)	0.12 (0.04–0.33)	0.35 (0.19–0.59)	0.16 (0.05–0.43)
Overall accuracy (95%CI)	0.71 (0.64–0.76)	0.64 (0.58–0.67)	0.67 (0.61–0.72)	0.58 (0.53–0.61)

95%CI, 95% confidence intervals; MDE, major depressive episode.

Receiver Operating Characteristics (ROC) analysis of the cut-off point (threshold) of summary score of PHQ-9 against “current major depressive episode”

Figure 1 displays the ROC curve of the cut-off point (threshold) of the summary score of PHQ-9 against “current major depressive episode”. The area under the curve (AUC) of 0.79 shows that the PHQ-9 summary scores can discriminate moderately between subjects with and without current major depressive episode [11]. Maximal discrimination between the presence and absence of “current major depressive episode” is achieved at the cut-off point which has the highest sum of sensitivity and specificity, as indicated by the “Youden index”.



The optimal cut-off was 13/14 (sensitivity 0.86, specificity 0.67) in the setting of a psychiatric specialty clinic and higher than that (9/10) recommended in primary care facilities and other specialty clinics [3,4,6]. Screening purposes demand high sensitivity and negative predictive value. Such was the case for a cut-off score of 11/12 or lower (Table 3). For diagnostic purposes, high specificity and positive predictive value are necessary. However, this condition was not reached with any cut-off score because specificity was higher than 90 % with a cut-off score of 21/22 or higher, but positive predictive values of all cut-off scores were lower than 65 %.

Correlation between PHQ-9 summary scores and HDRS, MADRS, or GAF

In patients with major depressive disorder, PHQ-9 summary scores were positively correlated with the severity measure of the depressive symptoms, HDRS and MADRS scores, but were negatively correlated with the overall levels of psychological, social, and occupational functioning, GAF scores [HDRS, $r = 0.55$, 95%CI (0.30, 0.73), $P = 0.0002$, $N = 42$; MADRS, $r = 0.68$, 95%CI (0.45, 0.82), $P < 0.0001$, $N = 36$; GAF, $r = -0.59$, 95%CI (-0.78, -0.30), $P = 0.0005$, $N = 31$].

Discussion

This study revealed that PHQ-9 has high sensitivity and a high negative predictive value in the setting of a clinic specializing in psychiatry as well as in primary care facilities and other specialty clinics [1,3,4,6] so that PHQ-9 is useful for screening purposes for the presence of a current major depressive episode. However, low specificity and a low positive predictive value in the setting of a clinic specializing in psychiatry do not support the use

Table 3 Sensitivity, specificity, positive and negative predictive values (PPV and NPV) at different cut-off scores for PHQ-9 summary scores against "current major depressive episode"

Cut-off	Sensitivity	Specificity	PPV	NPV
7/8	0.94	0.38	0.42	0.93
8/9	0.94	0.45	0.45	0.94
9/10	0.94	0.50	0.47	0.94
10/11	0.92	0.52	0.48	0.93
11/12	0.90	0.56	0.50	0.92
12/13	0.88	0.62	0.53	0.91
13/14 *	0.86	0.67	0.56	0.91
14/15	0.76	0.72	0.57	0.86
15/16	0.66	0.76	0.57	0.82
16/17	0.60	0.80	0.59	0.80
17/18	0.56	0.82	0.60	0.79
18/19	0.48	0.87	0.65	0.78
19/20	0.40	0.88	0.63	0.75
20/21	0.30	0.89	0.58	0.72
21/22	0.20	0.94	0.62	0.71
22/23	0.14	0.95	0.58	0.70

* Denotes the maximum sum of sensitivity and specificity.

of PHQ-9 for diagnostic purposes in contrast to the setting of primary care facilities and other specialty clinics, for which high specificity and a high positive predictive value were reported [1,3,4,6,8]. These findings were shown similarly for both the diagnostic algorithmic threshold and the summary score threshold (PHQ \geq 10), which were recommended for diagnostic and screening purposes by earlier studies [1,6].

Because of low specificity in both the diagnostic algorithmic threshold and the summary score threshold (PHQ \geq 10), false-positive cases should be noted in the use of PHQ-9. For a current major depressive episode, schizophrenia, panic disorder, adjustment disorder, eating disorders, dementia, and insomnia were diagnoses of false-positive cases. Compared with patients visiting primary care and other specialty clinics, more patients with various psychiatric disorders showing depressed mood and other various symptoms visit clinics specializing in psychiatry, which engenders more false-positive results.

PHQ-9 is used primarily for screening for the presence of major depressive episode but not major depressive disorder because the *DSM-IV-TR* diagnosis of major depressive disorder demands several exclusion criteria such as the absence of manic or hypomanic episode, but PHQ-9 does not include such exclusion items [4]. A significant number of bipolar disorder patients are invariably misdiagnosed with major depressive disorder by PHQ-9 because a major depressive episode is part of

bipolar disorder if one uses PHQ-9 for the screening for major depressive disorder. Originally, Kroenke et al. noted that before making a final diagnosis, the clinician is expected to rule out physical causes of depression, normal bereavement, and history of a manic episode [6]. Especially for a psychiatric specialty clinic, where bipolar disorder is much more prevalent, Kroenke's notion must be considered. For this reason, we compared the operational characteristics of PHQ-9 against "current major depressive episode" and "current major depressive episode with major depressive disorder". Our analysis of the validity of PHQ-9 for the screening for current major depressive episode with major depressive disorder (Table 2) revealed that the positive predictive value decreased by about 10 % compared with the screening for current major depressive episode. The underdiagnosis of bipolar disorder by PHQ-9 was a main reason for the increased false positives. Further diagnostic workup for past manic or hypomanic episodes or the combination of other screening tools for these episodes can resolve this major disadvantage of PHQ-9.

One might expect that the diagnostic algorithm threshold has better specificity than that of the summary score threshold because the diagnostic algorithm closely mimics the *DSM-IV-TR* diagnosis criteria of major depressive episodes. Nevertheless, against our expectations, the results obtained in this study and previous studies [3,4] showed no marked difference between two thresholds in operational characteristics. The summary score threshold (PHQ \geq 10) has slightly higher sensitivity and negative predictive value, but slightly lower specificity and positive predictive value than the diagnostic algorithm with no marked difference in this study and a previous study of primary care [4]. The ROC analysis of the cut-off point (threshold) of summary score of PHQ-9 against "current major depressive episode" showed that the optimal cut-off was 13/14, which showed 0.86 of sensitivity and 0.67 of specificity comparable to those of the diagnostic algorithm, in the setting of a psychiatric specialty clinic. Table 3 shows that high specificity (>90 %) was reached with a cut-off score of 21/22 or higher, but a high positive predictive value was not reached with any cut-off score. The salient implication is that PHQ-9 used in a psychiatric specialty clinic might be suitable for screening purposes with the optimal cut-off of 13/14 of the summary scores for major depressive episode, but not for diagnostic purposes. The summary score threshold with different cut-off points for specific purposes might be preferred to the diagnostic algorithm.

Summary scores of PHQ-9 in patients with major depressive disorder were moderately correlated with severity measures of the depressive symptoms, HDRS and MADRS scores, positively, and with the overall levels of psychological, social, and occupational functioning, GAF

scores, negatively. Consistent with our results, an earlier report of primary care described that PHQ-9 scores were correlated linearly with measures of quality of life, self-reported disability days, clinical visits, and self-reported difficulties related to symptoms [6]. In depressive disorder patients of primary care facilities, PHQ-9 scores were correlated moderately with the HDRS (17 items) scores [13]. In this study, the correlation of PHQ-9 scores was highest with the MADRS, which is related to the core concept of depression and which showed about twice the precision in estimating depression as the HDRS (17 items) showed for the average severity of depression [14]. Therefore, PHQ-9 scores might reflect the core symptoms of major depressive disorder, as inferred from items corresponding to the *DSM-IV* criteria items. To date, the correlation of PHQ-9 scores with the standard rating scales of depression, especially the MADRS, has not been reported in psychiatric patients. In addition to screening, PHQ-9 might be useful for measuring the severity of major depressive disorder.

All subjects in this study were psychiatric patients of a university hospital that provides primary and secondary services in Japan. These patients might have more complicated backgrounds than patients in other psychiatric clinics. Accordingly, these findings might not be generally applicable to other populations, which constitutes one limitation of this study.

Conclusions

PHQ-9, a brief questionnaire, reportedly has excellent reliability and validity for screening of major depressive episode and for measuring the severity of depressive symptoms [2]. It has been used in primary care and medical specialty clinics. Results of this study suggest that PHQ-9 is useful also in clinics specializing in psychiatry. This report is the first of a study assessing the validity of the use of PHQ-9 in such clinics. However, caution is necessary for the use of PHQ-9 in psychiatry because PHQ-9 is useful only for screening purposes for "current major depressive episode" as a result of its low positive predictive value. The cut-off point 13/14 of PHQ-9 summary scores, which is higher than that (≥ 10) recommended in primary care, was optimal [6].

Competing interests

The authors report no financial or other relationship that is relevant to the subject of this article. TK has received honoraria from GlaxoSmithKline, Astellas, and Eli Lilly, has received research/grant support from Astellas and GlaxoSmithKline, and is a member of the advisory boards of GlaxoSmithKline and Mitsubishi Tanabe Pharma. The other authors declare that they have no actual or potential conflict of interest.

Authors' contributions

TI designed the study and wrote the protocol. All other authors followed up and evaluated patients, and checked the protocol and discussion. All authors contributed to and approved the final manuscript.

Received: 6 November 2011 Accepted: 3 July 2012
Published: 3 July 2012

References

1. Spitzer RL, Kroenke K, Williams JB, PHQ Primary Care Study Group: Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. *JAMA* 1999, **282**:1737-1744.
2. Furukawa TA: Assessment of mood: guides for clinicians. *J Psychosom Res* 2010, **68**:581-589.
3. Gilbody S, Richards D, Brealey S, Hewitt C: Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): a diagnostic meta-analysis. *J Gen Intern Med* 2007, **22**:1596-1602.
4. Wittkampf KA, Naeije L, Schene AH, Huyser J, van Weert HC: Diagnostic accuracy of the mood module of the Patient Health Questionnaire: a systematic review. *Gen Hosp Psychiatry* 2007, **29**:388-395.
5. Löwe B, Unützer J, Callahan CM, Perkins AJ, Kroenke K: Monitoring depression treatment outcomes with the patient health questionnaire-9. *Med Care* 2004, **42**:1194-1201.
6. Kroenke K, Spitzer RL, Williams JB: The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med* 2001, **16**:606-613.
7. American Psychiatric Association: *Diagnostic and statistical manual of mental disorders: DSM-IV-TR*. 4th edition. Washington, DC: American Psychiatric Association; 2000.
8. Muramatsu K, Miyaoka H, Kamijima K, Muramatsu Y, Yoshida M, Otsubo T, Gejyo F: The patient health questionnaire, Japanese version: validity according to the mini-international neuropsychiatric interview-plus. *Psychol Rep* 2007, **101**:952-960.
9. Montgomery SA, Åsberg M: A new depression scale designed to be sensitive to change. *Br J Psychiatry* 1979, **134**:382-389.
10. Williams JB: A structured interview guide for the Hamilton depression rating scale. *Arch Gen Psychiatry* 1988, **45**:742-747.
11. Akobeng AK: Understanding diagnostic tests 3: Receiver operating characteristic curves. *Acta Paediatr* 2007, **96**:644-647.
12. Perlis RH, Brown E, Baker RW, Nierenberg AA: Clinical features of bipolar depression versus major depressive disorder in large multicenter trials. *Am J Psychiatry* 2006, **163**:225-231.
13. Wittkampf K, van Ravesteijn H, Baas K, van de Hoogen H, Schene A, Bindels P, Lucassen P, van de Lisdonk E, van Weert H: The accuracy of Patient Health Questionnaire-9 in detecting depression and measuring depression severity in high-risk groups in primary care. *Gen Hosp Psychiatry* 2009, **31**:451-459.
14. Carmody TJ, Rush AJ, Bernstein I, Warden D, Brannan S, Burnham D, Woo A, Trivedi MH: The Montgomery Asberg and the Hamilton ratings of depression: a comparison of measures. *Eur Neuropsychopharmacol* 2006, **16**:601-611.

doi:10.1186/1471-244X-12-73

Cite this article as: Inoue et al.: Utility and limitations of PHQ-9 in a clinic specializing in psychiatric care. *BMC Psychiatry* 2012 **12**:73.

Submit your next manuscript to BioMed Central
and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit



