Psychometric Properties of an Automated Telephone-Based PHQ-9

Ramesh Farzanfar, PhD,¹ Timothy Hereen, PhD,²
Joseph Fava, PhD,³ Jillian Davis, MPH,¹
Louis Vachon, MD,⁴ and Robert Friedman, MD¹

¹Medical Information Systems Unit, Boston University Medical Center, Boston, Massachusetts.

²Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts.

³Centers for Behavioral and Preventive Medicine, The Miriam Hospital, Providence, Rhode Island.

⁴Department of Psychiatry, Boston University School of Medicine, Boston, Massachusetts.

Abstract

Objective: This study aims to evaluate the psychometric properties of an automated version of the Patient Health Questionnaire-9 (PHQ-9) to further facilitate its use in primary care. We automated the PHQ-9 using a computer telephony modality (interactive voice response system) called telephone-linked communication (TLC). **Subjects and Methods:** Eighty subjects were divided into four depression categories: none, mild, moderate, and severe. The automated questionnaire, TLC-PHQ-9, was administered to all subjects five times over a 3-month period, at intervals of 0 (T1), 1, 3, 4, and 4 (T5) weeks, respectively. The Inventory of Depressive Symptomatology (IDS) was administered (paper-and-pencil) at T1 and T5. We examined (1) reliability, (2) validity, (3) sensitivity and specificity, and (4) sensitivity to change. **Results:** (1) Test-retest reliability showed substantial agreement between T1 and T2, with a weighted kappa of 0.76 (95% confidence interval [CI] 0.67–0.85). Cronbach's coefficient alpha values ranged from 0.913 to 0.918 for each TLC-PHQ-9 assessment. (2) The weighted kappa of 0.78 (95% CI 0.70–0.87) for T1 and 0.73 (95% CI 0.63–0.83) for T5 showed strong agreement between TLC-PHQ-9 and IDS in all depression categories. (3) TLC-PHQ-9 demonstrated good sensitivity (82.4%) and very good specificity (90.7%) for moderate-plus depression and poorer sensitivity (54.2%) but very good specificity (97.8%) for severe-plus depression. (4) The weighted kappa of 0.53 (95% CI 0.35–0.70) indicated moderate agreement between TLC-PHQ-9 and IDS. **Conclusions:** An automated telephony administration of the PHQ-9 appears to be a valid and reliable tool for monitoring depression symptoms and has strong fidelity across patients.

Key words: home health monitoring, telehealth, e-health

Introduction

Assessment of depression and its treatment is increasingly being carried out by primary care physicians. Identification of depression at the primary care level typically involves utilization of standardized validated tools. However, many depression assessment questionnaires are long and time-consuming and thus burdensome for both patients and clinicians. The Patient Health Questionnaire-9 (PHQ-9) (see *Appendix*), a brief depression assessment instrument with severity categories, has shown good reliability and validity¹ and has gained recognition and increased use in primary care. The PHQ-9 is succinct and thus feasible for administration in a demanding clinical environment. The instrument has also proven helpful in assessment of depression among patients with chronic diseases and can be utilized by primary care physicians successfully, thus enabling clinicians to better manage patients' depression symptoms.^{2,3}

The aim of this study is to evaluate the psychometric properties of an automated version of the PHQ-9 to further facilitate the use of the questionnaire in primary care. We automated the PHQ-9 using a computer telephony modality (interactive voice response [IVR] system) called telephone-linked communication (TLC). We subsequently evaluated the TLC-PHQ-9 among 80 participants for test-retest reliability, validity (using the Inventory of Depressive Symptomatology [IDS] as the gold standard), sensitivity, specificity, and sensitivity to change. Results demonstrated that TLC-PHQ-9 has good psychometric properties and is an efficient screening and monitoring tool that can be reliably used in primary care.

DEPRESSION AND ITS IMPACT

Depression is the leading cause of disability in the United States⁴ and the third most important cause of disease burden worldwide.⁵ In the United States, a national face-to-face household survey showed that 6.7% of adults experienced a major depressive episode in the past year.⁴ Women demonstrate a higher prevalence of depression (11.7%) than men (5.6%),⁶ and Hispanics (5.17%) and blacks (4.57%) show lower rates than whites (6.52%).⁷

Research shows that 51% of those with major depression suffer from lifetime anxiety⁸ and that depression usually co-occurs with anxiety, stress,⁹ and posttraumatic stress disorder.¹⁰ Furthermore, there is an increased rate of alcohol and drug abuse among depressed individuals.¹¹ Depression sufferers not only have a higher rate of dying¹² from risk factors such as cardiovascular disease,^{13,14} they may also have a shorter life span because depression is the most common precursor to suicidality.¹⁵

Moreover, depression not only threatens marital and family stability, it also strains relationships with friends, neighbors, and other acquaintances, potentially leading to disintegration of relationships and sometimes dissolution of family structure.¹⁶ It is further documented that patients experience significant impairment in their social roles such as difficulty functioning in the workplace, school, and other social settings.¹⁷ This creates an immense economic burden on the individual patient and for the community at large.¹⁸ The urgency for identification of depression symptoms at an early stage is compounded by the fact that, if not treated, depression is likely to become chronic as only one episode increases the risk for subsequent events to 50%.¹⁹

PHQ-9: A FEASIBLE TOOL TO SCREEN AND MONITOR FOR DEPRESSION

The PHQ-9 is the depression module of the Primary Care Evaluation of Mental Disorders^{20,21} that was designed to be used in primary care²² and provides scores on each of the nine Diagnostic and Statistical Manual of Mental Disorders, 4th edition (DSM-IV) criteria using a severity scale from "0" (not at all) to "3" (nearly every day). The advantages of the PHQ-9 include self-administration or administration by a clinician, brevity, the efficient follow-up of depression symptom severity over time, and acceptable reliability and validity.^{20,22}

Administration of the PHQ-9 to 6,000 patients across eight primary care and seven obstetrics-gynecology clinics produced a sensitivity of 88% and a specificity of 88% for major depression, using the mental health professional interview as the criterion standard. Cronbach's coefficient alpha (CCA) internal consistency reliability coefficients ranged from 0.86 to 0.89, whereas 2-day test-retest reliability was estimated to be 0.84 with a nearly identical mean total score. PHQ-9 scores were also strongly correlated with the Mental Health Inventory (five-item mental health scale of the Short Form-20).¹ The PHQ-9 has been evaluated for test-retest reliability and validity over the telephone in a study of 346 participants, comparing self-administration and telephone administration. The results demonstrated a high correlation coefficient ($r=0.82$) and weighted kappa ($\kappa > 0.58$) between telephone administration and self-administration. The telephone administration had a high internal consistency (CCA=0.82), which was comparable to the self-administered PHQ-9 (CCA=0.86).²³

Also, an automated telephone-based administration of the PHQ-9 using IVR compared with a paper-and-pencil administration among 51 Veterans showed similar results between the two forms of administration, with the CCA statistic for IVR administration (CCA=0.76) comparable to that for pencil-and-paper administration (CCA=0.82), with an intraclass correlation coefficient for the two modes of 0.65. The study also demonstrated that the automated administration is not as sensitive to high levels of depressive symptom severity.²⁴ The present study is unique in that it explores certain properties of an automated PHQ-9 that have not been previously evaluated.

IVR

An IVR system makes it possible for people to use a telephone to interact with a computer by pressing the keys on the telephone keypad. The genesis of this technology goes back to the 1930s before

digital computers became household items. In the 1960s, a tone dialing technology was developed at Bell Laboratories. Over time the technology was refined, and by the 1990s the technology was not only cost-effective but also sufficiently sophisticated to handle multiple calls and complex interactions. Modern advances have enabled IVR systems to collect users' responses through speech using speech recognition software.²⁵

An IVR platform includes hardware (servers, mainframes, and a telephony infrastructure) and software (e.g., a programming language such as Voice Extensible Markup Language and others) on which an IVR application runs. IVR systems can run a gamut of different applications that serve various purposes, including commercial endeavors. In medicine, IVR applications are increasingly playing important functions, including collection and retrieval of clinical data.

TLC-PHQ-9: AN IVR-BASED DEPRESSION SCREENING MODALITY

The TLC system has been developed by researchers at Boston University.²⁶⁻²⁸ TLC carries out automated telephone conversations with patients and can be programmed to screen patients for symptoms and/or deliver education and behavioral counseling for targeted self-care behaviors such as medication-taking, exercise promotion, smoking cessation, management of chronic disease, etc.²⁶⁻²⁸ Different applications of TLC have demonstrated the technology's ability to control and prevent exacerbations that can lead to morbidity, mortality, and preventable emergency department visits and hospitalizations.²⁶⁻³⁴

In this study, we developed an algorithm of the PHQ-9 questions that was subsequently programmed into the IVR platform using a recorded human voice. TLC stored the information collected from the study participants in a database, which was accessible by researchers for analysis following study completion.

Subjects and Methods

STUDY OBJECTIVES

The principal objectives of this research were to determine (1) test-retest reliability, (2) validity, (3) sensitivity and specificity, and (4) sensitivity to change of TLC-PHQ-9. The IDS was used as the gold standard.

STUDY PARTICIPANTS

The study was approved by the Boston University Institutional Review Board. In total, 89 study participants were enrolled at baseline, of whom 80 (89.9%) successfully completed each of the five post-screener assessments. The 80 individuals were divided into four categories of depression symptomatology: no depression and mild, moderate, and severe depression. Those who were eligible to participate were 18 years of age or older and fluent in English; the 20 nondepressed participants were enrolled based on these two criteria, whereas the rest had to have a current diagnosis of major depressive disorder (verified by contacting the patients' responsible clinician). All participants, including the nondepressed group, were screened for

depression symptoms using the PHQ-9. Depression group distribution by IDS score at baseline, along with baseline demographics, is shown in Table 1.

It should be pointed out that the TLC system was programmed to create an alert in case a study participant responded positively to Question 9 ("Have you had thoughts that you would be better off dead, or of hurting yourself in some way?"). Alerts were immediately sent to the study's Principal Investigator via e-mail and text messages. Subsequently, the study's in-house clinician, a psychiatrist, was contacted and made aware of the alert's content, including the name and contact information of the study participant. The study clinician would then contact the participant and make appropriate clinical decisions based on her conversation with the person.

STUDY PROCEDURES

The TLC-PHQ-9 was administered to all study participants upon enrollment for a total of five times over a 3-month period. The time interval between the five evaluation points was 1, 3, 4, and 4 weeks, respectively.

Table 1. Participant Demographics by Depression Group (n=80 Completers)

	TOTAL (N=80) [% (M)]
Gender	
Female	72.5 (58)
Age (years) [mean (SD)]	38.4 (14.1)
Married/living with partner	
No	63.8 (51)
Education	
Bachelor's degree or higher	63.8 (51)
Hispanic	
No	88.8 (71)
Ethnicity	
White	71.3 (57)
First diagnosis of depression	
No	68.8 (55)
Ever taken depression medications	
No	52.5 (42)
IDS depression category (Time 1)	
None	24 (30.0)
Mild	14 (17.5)
Moderate	26 (32.5)
Severe	16 (20.0)

IDS, Inventory of Depressive Symptomatology; SD, standard deviation.

During the first study visit (baseline [T1]) informed consent was obtained. Subsequently, participants completed the first administration of the TLC-PHQ-9. Moreover, all participants completed a self-administered, paper-and-pencil version of the IDS. The IDS is a 30-item instrument (developed by Rush et al.,³⁵⁻³⁹ with good psychometric properties) that was used as a gold standard to evaluate the validity of TLC-PHQ-9. The instrument includes all DSM-IV criterion items required to diagnose a major depressive episode and assess symptom severity. The second interaction with the TLC-PHQ-9 occurred 1 week later. The third contact was identical to the second but occurred 3 weeks later. Four weeks later study participants received the TLC-PHQ-9 for the fourth time. At the end of the fourth call, participants were scheduled to return to the study headquarters (in 4 weeks) (T5) for the final administration of the TLC-PHQ-9 and paper-and-pencil administration of the IDS.

The follow-up assessments were useful to detect change in depressive symptomatology. In addition, the first and last data collections provided an evaluation of the validity of TLC-PHQ-9 through comparison with a paper-and-pencil administration of the IDS.³⁵ The first and last data collections were also used to assess the ability of TLC-PHQ-9 to detect final change in the severity of depression (as compared against the IDS for validation purposes).

STATISTICAL ANALYSES

Analyses were conducted using IBM SPSS Statistics for Windows, release 20.0.0⁴⁰ and SAS version 9.3 for Windows.⁴¹ Principal component analysis (PCA) was used to examine the hypothesized internal structure of the PHQ-9 and the IDS at each assessment, whereas the CCA statistic was calculated to provide a measure of the internal consistency of each instrument. The weighted kappa was used to compare change in depression categorization between assessments. Pearson's correlation coefficient was calculated between the PHQ-9 and the IDS at T1 and T5, and a sensitivity and specificity analysis of the PHQ-9 was conducted using the IDS as the gold standard measure.

Weighted kappas were calculated to evaluate (1) test-retest reliability, (2) validity, and (3) sensitivity to change of the automated PHQ-9. Internal validity was assessed using PCA, and internal consistency was assessed using CCA. Sensitivity and specificity of the PHQ-9 as a screening test for depression were calculated using the IDS as the gold standard measure of depression. We performed two separate analyses: (1) using the PHQ-9 as a screener for moderate or higher depression (e.g., dichotomizing the IDS as none/mild versus moderate/severe/very severe) and (2) using the PHQ-9 as a screener for severe depression (e.g., dichotomizing the IDS as none/mild/moderate versus severe/very severe). For the primary analysis, the data were pooled from T1 and T5, and thus this analysis is based on n=160 depression categorizations from n=80 subjects. Separate analyses were also performed for T1 and T5. In this analysis, 95% confidence intervals (CIs) for sensitivity and specificity are given, using CI formulas that use a small-sample adjustment in calculating the standard error. Receiver operating characteristics curve analysis and the area under the curve treat the PHQ-9 screener as a 4-point

scale (plotting sensitivity by [1–specificity] for all possible cut-points on the PHQ-9 scale) and describe how well the scale predicts IDS depression. AUC values above 0.80 represent very good prediction, with values above 0.90 representing excellent prediction.

Results

TEST–RETEST RELIABILITY

The weighted kappa (Table 2), measuring the percentage of agreement beyond chance in these data, was 0.76 with a 95% CI (0.67–0.85), which indicates substantial agreement.⁴²

PHQ-9 internal structural validity and consistency. The PHQ-9 uses a single score based on the sum of the individual items in these instruments to measure severity of depression symptomatology. We examined the internal structure of the PHQ-9 using PCA to confirm the appropriateness of this scoring methodology. We also explored the magnitude of the item loadings for a one-dimensional solution at each administration of the PHQ-9. We expected each of the individual item loadings on the single component of the PCA solutions to be generally high (above 0.4, and preferably above 0.6).^{43,44} For the PHQ-9 this was the case for each of the six administrations (screener and five assessments) of the instrument. Item loadings for items 1–8 were generally above 0.6, and even the lowest loading item (number 9) was above 0.52 (Table 3). CCA was very high at each assessment. Each administration produced very similar CCA values that ranged from 0.913 to 0.918 (Table 4).

VALIDITY

Tables 4 and 5 demonstrate the agreement between the IDS and TLC-PHQ-9 for individuals in all symptom categories for T1 and T5, respectively.

The weighted kappa for T1 (Table 4) was 0.78 (95% CI 0.70–0.87), showing substantial agreement between the TLC-PHQ-9 and the IDS with respect to all symptom categories. The weighted kappa for T5 (Table 5) was 0.73 (95% CI 0.63–0.83), again showing substantial agreement between the PHQ-9 and the IDS with respect to all symptom categories.

Table 2. Test–Retest Reliability: Comparing Results of the Telephone-Linked Communication–Patient Health Questionnaire-9 at Time 1 and Time 2

TLC-PHQ-9 RATING AT TIME 1	TLC-PHQ-9 RATING AT TIME 2			
	NONE	MILD	MODERATE	SEVERE
None	24	2	0	0
Mild	3	7	3	0
Moderate	0	11	19	1
Severe	0	0	2	8

Data are number of subjects in each group.

PHQ-9, Patient Health Questionnaire-9; TLC, telephone-linked communication.

Table 3. Principal Component Item Loadings and Cronbach's Coefficient Alpha Reliability Values for the Telephone-Linked Communication–Patient Health Questionnaire-9 at Each Assessment

TLC-PHQ-9 ITEM	TIME 1	TIME 2	TIME 3	TIME 4	TIME 5
PHQ 1	0.917	0.871	0.89	0.878	0.888
PHQ 2	0.901	0.888	0.907	0.907	0.858
PHQ 3	0.796	0.824	0.772	0.837	0.803
PHQ 4	0.78	0.778	0.861	0.833	0.81
PHQ 5	0.658	0.734	0.752	0.647	0.677
PHQ 6	0.784	0.845	0.857	0.827	0.8
PHQ 7	0.839	0.798	0.777	0.808	0.816
PHQ 8	0.672	0.665	0.579	0.608	0.657
PHQ 9	0.579	0.576	0.52	0.558	0.57
Coefficient alpha	0.915	0.918	0.918	0.915	0.913

PHQ-9, Patient Health Questionnaire-9; TLC, telephone-linked communication.

SENSITIVITY AND SPECIFICITY

As a screener for moderate-plus depression, the PHQ-9 has good sensitivity (82.4%) and very good specificity (90.7%). A sensitivity of 82.4% indicates that, for those with moderate-or-higher depression (IDS), 82.4% will screen positive on the PHQ-9. From the specificity result, 90.7% of those with none/mild depression will screen negative on the PHQ-9. The area under the curve for the PHQ-9 in predicting moderate or higher depression is 0.925, indicating excellent prediction.

As a screener for severe-plus depression, the PHQ-9 has poorer sensitivity (54.2%) but very good specificity (97.8%). This indicates that, among those who truly have severe depression (IDS), only 54.2% will screen positive for severe depression using PHQ-9.

Table 4. Comparing Results of the Telephone-Linked Communication–Patient Health Questionnaire-9 and Inventory of Depressive Symptomatology at Time 1

TLC-PHQ-9 RATING AT TIME 1	IDS RATING AT TIME 1			
	NONE	MILD	MODERATE	SEVERE
None	24	2	0	0
Mild	0	7	6	0
Moderate	0	5	19	7
Severe	0	0	1	9

Data are number of subjects in each group.

IDS, Inventory of Depressive Symptomatology; PHQ-9, Patient Health Questionnaire-9; TLC, telephone-linked communication.

Table 5. Comparing Results of the Telephone-Linked Communication-Patient Health Questionnaire-9 and Inventory of Depressive Symptomatology at Time 5

TLC-PHQ-9 RATING AT TIME 5	IDS RATING AT TIME 5			
	NONE	MILD	MODERATE	SEVERE
None	27	7	1	0
Mild	0	11	6	0
Moderate	0	2	16	4
Severe	0	1	1	4

Data are number of subjects in each group.

IDS, Inventory of Depressive Symptomatology; PHQ-9, Patient Health Questionnaire-9; TLC, telephone-linked communication.

However, of those without severe depression, 97.8% will screen as not having severe depression on the PHQ-9. The area under the curve for the PHQ-9 in predicting severe depression is 0.913, again indicating excellent prediction.

SENSITIVITY TO CHANGE

We also explored the capacity of the PHQ-9 to detect change in a patient's depression symptom severity over time. TLC-PHQ-9 was administered five times to the study participants during the 3-month period of their participation. *Table 6* shows corresponding change on the TLC-PHQ-9 and IDS from T1 to T5. The weighted kappa of 0.53 (95% CI 0.35–0.70) indicates that there is moderate agreement. The TLC-PHQ-9 was somewhat less likely to show improvement than the IDS. In fact, of the 23 patients showing improvement on the IDS, 8 (35%) showed no change on the TLC-PHQ-9.

Table 6 also demonstrates a trend toward improvement among the study participants. As noted, the number of individuals with no depression as well as mild depression symptoms improved in both PHQ-9 and IDS scores.

Table 6. Sensitivity to Change from Time 1 to Time 5: Telephone-Linked Communication-Patient Health Questionnaire-9 and Inventory of Depressive Symptomatology

CHANGE IN TLC-PHQ-9 CATEGORY	CHANGE IN IDS CATEGORY		
	IMPROVEMENT	NO CHANGE	WORSENING
Improvement	15	2	0
No change	8	36	3
Worsening	0	5	1

Data are number of subjects in each group.

IDS, Inventory of Depressive Symptomatology; PHQ-9, Patient Health Questionnaire-9; TLC, telephone-linked communication.

Discussion

This research shows that an automated administration of the PHQ-9 is a valid and reliable tool for monitoring depression symptoms. The internal structural validity of the TLC-PHQ-9 is strongly supported by the high PCA item loadings and excellent CCA values. In addition, the internal structural validity of the IDS is also supported, although a few items had low loadings. The CCA values for the IDS were excellent, however, and demonstrate the advantage of increased reliability that a long instrument can attain as it counterbalanced the few poorly performing items on the scale.

The number of participants in each depression group as categorized by the TLC-PHQ-9 changed across time, with the size of the nondepressed group increasing and the size of the severe depression group decreasing. This observation was mirrored with the IDS (*Tables 4–6*), thus revealing a strong concordance between the TLC-PHQ-9 and our gold standard, the IDS.

The sensitivity and specificity analysis compared the two depression questionnaires and suggests that the TLC-PHQ-9 performs very well as a screener for moderate or higher depression but only moderately well as a screener for severe depression. Therefore, the system can be used to effectively monitor patients who suffer mild-moderate depression in a primary care environment. However, patients with severe depression symptoms need closer supervision by a responsible clinician, and thus it is not advisable for them to be monitored by an automated system.

With respect to sensitivity to change, although test-retest scores for participants who receive treatment are not expected to fluctuate too widely over time, some improvement in scores is expected. We only observed moderate concordance between the TLC-PHQ-9 and the IDS regarding sensitivity to change. A larger sample is required in order to establish the meaningfulness of the automated version of the TLC-PHQ-9 to detect change.

With regard to acceptability of the system, we believe the system was acceptable to patients. In fact, although the study did not include a formal usability/acceptability protocol, the study participants had the option of leaving a message for the study staff describing their experience using the TLC-PHQ-9. Approximately one-third of the study participants ($n = 31$) chose to leave a message that was recorded by the TLC. We found the messages positive and categorized them as follows: (1) enjoying the “activity” of engagement with TLC; (2) having a private and anonymous listener (TLC) with which to share one's innermost feelings; and (3) reflecting on the responses to the questions posed by TLC and thus learning more about one's symptom status. Finally, automation of the PHQ-9 facilitates self-monitoring and self-report that might enhance help-seeking behavior by patients based on the theories of mere-measurement and/or question-to-behavior effects.^{45–49}

As for primary care clinicians, automated administration of the PHQ-9 might be helpful to those who manage patients with mild-moderate depression symptoms. The fidelity of administration is high, and there will be time saved that could be devoted to other aspects of the clinical visit. Furthermore, an automated telephone-based version of PHQ-9 can be programmed to alert clinicians of changes in their patients' depression symptoms over time.

Acknowledgments

This research was funded by the National Institute of Mental Health.

Disclosure Statement

No competing financial interests exist.

REFERENCES

- Kroenke K, Spitzer R, Williams J. The PHQ-9: Validity of a brief depression severity measure. *J Gen Intern Med* 2001;16:606–613.
- Chen TM, Huang FY, Chang C, Chung H. Using the PHQ-9 for depression screening and treatment monitoring for Chinese Americans in primary care. *Psychiatr Serv* 2006;57:976–981.
- McFeature B, Pierce TW. Primary care behavioral health consultation reduces depression levels among mood-disordered patients. *J Health Disparities Res Pract* 2012;5:36–44.
- Kessler RC, Chiu WT, Demler O. Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Arch Gen Psychiatry* 2005;62:617–627.
- World Health Organization. *The global burden of disease: 2004 update*. Geneva: WHO Press, 2008.
- Ford DE, Erlinger TP. Depression and C-reactive protein in US adults: Data from the Third National Health and Nutrition Survey. *Arch Intern Med* 2004;164:1010–1014.
- Oquendo MA, Lizardi D, Greenwald S. Rates of lifetime suicide attempt and rates of lifetime major depression in different ethnic groups in the United States. *Acta Psychiatr Scand* 2004;110:446–451.
- Kessler RC, Nelson C, McGonagle KA. Comorbidity of DSM-III-R major depressive disorder in the general population: Results from the US National Comorbidity Survey. *Br J Psychiatry* 1996;168(Suppl 30):17–30.
- Hirschfeld RMA. The comorbidity of major depression and anxiety disorders: Recognition and management in primary care. *Prim Care Companion J Clin Psychiatry* 2001;3:244–254.
- Shalev AY, Freedman S, Peri T, Brandes D, Sahar T, Orr SP, Pitman RK. Prospective study of posttraumatic stress disorder and depression following trauma. *Am J Psychiatry* 1998;155:630–637.
- Grant BF. Comorbidity between DSM-IV drug use disorders and major depression: Results of a national survey of adults. *J Subst Abuse* 1995;7:481–487.
- Rush AJ. The varied clinical presentations of major depressive disorder. *J Clin Psychiatry* 2007;68(Suppl 8):4–10.
- Alboni P, Favaron E, Paparella N, Sciammarella M, Pedaci M. Is there an association between depression and cardiovascular mortality or sudden death? *J Cardiovasc Med* 2008;9:356–362.
- Taylor WD, McQuoid DR, Ranga Raman Krishnan K. Medical comorbidity in late life depression. *Int J Geriatr Psychiatry* 2004;19:935–943.
- Cassano P, Fava M. Depression and public health: an overview. *J Psychosom Res* 2002;53:849–857.
- Freeman A, Epstein N, Simon KM. Introduction. In: Freeman A, Epstein N, Simon KM, eds. *Depression and the Family*. New York: The Haworth Press, 1986:5–7.
- Hirschfeld RMA, Montgomery SA, Keller MB, Kasper S, Schatzberg AF, Moller HJ, et al. Social functioning in depression: A review. *J Clin Psychiatry* 2000;61:268–275.
- Simon GE. Social and economic burden of mood disorders. *Biol Psychiatry* 2033;54:208–215.
- Burcasa SL, Lacono WG. Risk for recurrence in depression. *Clin Psychol Rev* 2007;27:959–985.
- Spitzer RL, Kroenke K, Williams JB. Validation and utility of a self report version of PRIME-MD: The PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire. *JAMA* 1999;282:1737–1744.
- Spitzer R, Williams JB, Kroenke K. Validity and utility of the PRIME-MD Patient Health Questionnaire in assessment of 3000 obstetric-gynecologic patients: The PRIME-MD Obstetric-Gynecology Study. *Am J Obstet Gynecol* 2000;183:759–769.
- Lowe B, Unutzer J, Callahan CM. Monitoring depression treatment outcomes with the Patient Health Questionnaire-9. *Med Care* 2004;42:1194–1201.
- Pinto-Meza A, Serrano-Blanco A, Penarrubia MT. Assessing depression in primary care with the PHQ-9: Can it be carried out over the telephone? *J Gen Intern Med* 2005;20:738–742.
- Turvey C, Sheeran T, Dindo L. Validity of the Patient Health Questionnaire, PHQ-9, administered through interactive-voice response technology. *J Telemed Telecare* 2012;18:348–351.
- Wikipedia. Interactive voice response. Available at http://en.wikipedia.org/wiki/Interactive_voice_response (last accessed June 3, 2013).
- Friedman R, Kazis LE, Jette A. A telecommunication system for monitoring and counseling patients with hypertension: Impact on medication adherence and blood pressure control. *Am J Hypertens* 1996;9:285–292.
- Friedman R, Stollerman JE, Mahoney DM, Rozenblyum L. The virtual visit: Using telecommunications technology to take care of patients. *J Am Med Inform Assoc* 1997;4:413–425.
- Friedman R, Stollerman JE, Rozenblyum L. A telecommunications system to manage patients with chronic disease. *Medinfo*. Seoul: International Medical Informatics Association, 1998:1330–1334.
- Farzanfar R, Stevens A, Vachon L, et al. Design and development of a workplace mental health assessment and intervention system. *J Med Syst* 2007;31:49–62.
- Farzanfar R, Stevens A, Pham Q, Friedman R. A formative qualitative evaluation of usability and acceptability of a workplace mental health assessment and intervention system. *Int J Ment Health Promot* 2008;10:17–25.
- Farzanfar R, Locke S, Heeren T, et al. Workplace tele-communications technology to identify mental health disorders and facilitate self-help or professional referrals. *Am J Health Promot* 2011;25:207–216.
- Farzanfar R, Finkelstein D. Evaluation of a workplace technology for mental health assessment: A meaning-making process. *Comput Hum Behav* 2012;28:160–165.
- Delichatsios H, Glanz K, Tennstedt S, et al. Randomized trial of a "talking computer" to improve adults' eating habits. *Am J Health Promot* 2001;15:215–224.
- Pinto B, Marcus BH, Kelley H. Effects of a computer-based, telephone-counseling system on physical activity. *Am J Prev Med* 2002;23:113–120.
- Rush AJ, Hiser W, Giles DE. A comparison of self-reported versus clinician-rated symptoms in depression. *J Clin Psychiatry* 1987;48:246–248.
- Rush AJ, Gullion CM, Basco MR. The Inventory of Depressive Symptomatology (IDS): Psychometric properties. *Psychol Med* 1996;26:477–486.
- Rush AJ, Trivedi MH, Ibrahim HM. The 16-item Quick Inventory of Depressive Symptomatology (QIDS) Clinician Rating (QIDS-C) and Self-Report (QIDS-SR): A psychometric evaluation in patients with chronic major depression. *Biol Psychiatry* 2003;54:573–583.
- Rush AJ, Trivedi MH, Carmody TJ. One-year clinical outcomes of depressed public-sector outpatients: A benchmark for subsequent studies. *Biol Psychiatry* 2004;56:46–53.
- Rush AJ, Trivedi MH, Carmody TJ. Self-reported depressive symptom measures: Sensitivity to detecting change in a randomized, controlled trial of chronically depressed, nonpsychotic outpatients. *Neuropsychopharmacology* 2005;30:405–416.
- IBM Corp. 2011. Available at www.ibm.com (last accessed May 23, 2013).
- SAS Institute Inc. 2002–2011. Available at www.sas.com (last accessed May 23, 2013).

42. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* **1977**;33:159–174.
43. Velicer WF, Fava JL. An evaluation of the effects of variable sampling on component, image, and factor analysis. *Multivar Behav Res* **1987**;22:193–209.
44. Velicer WF, Fava JL. The effects of variable and subject sampling on factor pattern recovery. *Psychol Methods* **1998**;3:231–251.
45. Fischer EH, Farina A. Attitudes toward seeking professional psychological help: A shortened form and considerations for research. *J Coll Student Dev* **1995**;36:368–373.
46. Godin G, Belanger-Gravel A, Amireault S. The effect of mere-measurement of cognitions on physical activity behavior: A randomized controlled trial among overweight and obese individuals. *Int J Behav Nutr Physiol* **2011**;8:2.
47. Jacomb PA, Jorm AF, Rodgers B. Emotional response of participants to a mental health survey. *Soc Psychiatry Psychiatr Epidemiol* **1999**;34:80–84.
48. Morwitz VG, Fitzsimons GJ. The mere-measurement effect: Why does measuring intentions change actual behavior. *J Consum Psychol* **2004**; 4:64–74.
49. Williams P, Block LG, Fitzsimons GJ. Simply asking questions about health behaviors increases both healthy and unhealthy behaviors. *Soc Influence* **2006**;2:117–127.

Address correspondence to:
Ramesh Farzanfar, PhD
Medical Information Systems Unit
Boston University Medical Center
801 Massachusetts Avenue, 2nd Floor
Boston, MA 02118

E-mail: rfarzanf@bu.edu

Received: April 30, 2013

Revised: June 6, 2013

Accepted: June 6, 2013

Appendix

Table A1. Patient Health Questionnaire-9

QUESTION	OVER THE LAST 2 WEEKS, HOW OFTEN HAVE YOU BEEN BOTHERED BY ANY OF THE FOLLOWING PROBLEMS?	NOT AT ALL	SEVERAL DAYS	MORE THAN HALF THE DAYS	NEARLY EVERY DAY
Q1	Little interest or pleasure in doing things	0	1	2	3
Q2	Feeling down, depressed, or hopeless	0	1	2	3
Q3	Trouble falling or staying asleep, or sleeping too much	0	1	2	3
Q4	Feeling tired or having little energy	0	1	2	3
Q5	Poor appetite or overeating	0	1	2	3
Q6	Feeling bad about yourself—or that you are a failure or have let yourself or your family down	0	1	2	3
Q7	Trouble concentrating on things, such as reading the newspaper or watching television	0	1	2	3
Q8	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	0	1	2	3
Q9	Thought that you would be better off dead, or of hurting yourself in some way	0	1	2	3

The Patient Health Questionnaire-9 is scored by adding up all checked boxes. Point scores as assigned as follows: not at all=0; several days=1; more than half the days=2; and nearly every day=3. A total score of 0–4 is defined as no depression, 5–9 as mild depression, 10–19 as moderate depression, and 20–27 as severe depression.