More Effective and Cost-Effective Use of the PHQ-9

Abstract

In this paper we analyze the implementation of the PHQ-9 for depression screening from the perspective of depth and breadth, accuracy and automated processing. Using Peek's three world view of clinical, operational and financial as a simplified framework, we show the high cost and limited effectiveness of current implementations and suggest ways to improve the process to help enhance and advance efforts to integrate mental health into primary care.

Introduction

Screening in primary care for the most common mental health disorders remains too low.[1] Rates of depression continue to rise, and globally, depression remains a leading cause of disability.[2] The United States Preventive Service Task Force (USPSTF) recommends screening adults for depression in primary care when the adequate supports are in place.[3] Taking into account the prevalence of mental health conditions in primary care, and the robust evidence, having less than 5% screening rates for depression in primary care is an area of opportunity for profound growth; you can't treat what you don't identify or diagnose.[4] Screening for depression does not have to be difficult, but it does require a conscious decision about workflow, delivery, and ultimately practice culture.[5,6]

Due to its ease-of-use and reasonable sensitivity and specificity, the PHQ-9 has become the de facto standard for depression screening in medical and mental health settings.[7] Using the PHQ-9 to detect and monitor psychological distress is considered to be standard practice, but unfortunately how this gets implemented often becomes costly and relatively ineffective.[8,9] To better address mental health in primary care, screening and treatment will be essential components that require a practice to pay attention to clinical, operational, and financial factors simultaneously.[10]

Operationalizing the PHQ-9

From the perspective of depth and breadth, depression screening with the PHQ-2 or PHQ-9 is similar to taking a patient's temperature. An elevated score indicates something is going on, but often clinicians have little recourse other than to refer on for further evaluation. If the screening automatically probed further for other symptoms or related issues such as mania, suicidality, anxiety, panic, trauma, or social factors as indicated, the depression screening would yield far more effective, clinically-actionable information.

Looking at accuracy, the PHQ-9 is reported to be 88% sensitive and 88% specific.[8,9] This unfortunately means 12% of the time a clinician is receiving false negatives, as well as 12% of the time false positives. False negatives create cost for the system and false positives waste time and create needless cost for the practice, let alone the impact it can have on a patient and their family

In this paper, we introduce probabilistic scoring as a higher accuracy alternative, which enables dynamic administration of the PHQ-9. A dynamic administration approach can help reduce the number of questions required, enabling more measures to be probed.

Finally, expanding depth and breadth, and improving accuracy using probabilistic scoring, requires electronic screening, which also enables automated processing. Using self-report measures, patients can do the data entry. Patient information can freely flow in and out of the electronic medical record (EMR), limiting or potentially eliminating staff involvement. Using the PHQ-9 is a good first step, but how you implement makes a significant difference.

Methods

Using practice-based research, over a 6 month period, a Federally Qualified Health Center (FQHC) based in Montana expanded its typical self-report PHQ-9 screenings administered on tablets, by automatically administering a Quick PsychoDiagnostic panel (QPD) computer assessment randomly before or after administering the PHQ-9. A study dataset of 2495 administrations of the combined PHQ-9 and QPD were collected.

The QPD is a computer assessment that follows the DSM-V, dynamically branching into additional questions as indicated by prior answers. It generates severity scores for depression, manic episode, anxiety, panic disorder, PTSD, eating disorder, substance use and somatization, and computes provisional diagnoses for major depression, dysthymia, depression NOS, bipolar, generalized anxiety, panic, OCD, anxiety NOS, anxiety secondary to depression, bulimia, substance abuse, somatization, suicidal ideation and suicidal risk.[13] Self-report administration of the QPD typically averages 6 ½ minutes per administration. The QPD is considered to have a positive result when any one of the severity scales exceeds its cut score.

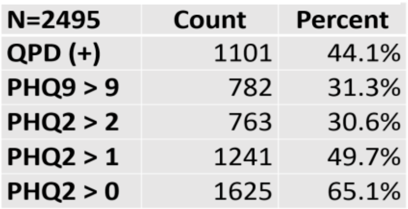
Using spreadsheet logic and calculations, the study dataset is analyzed for the number of false negative and false positive results based on different cut scores and the resulting costs if the QPD is used, rather than having a health professional gather the same information. This analysis also examines the information that is missed and could have been gathered if the PHQ-2 or PHQ-9 had been used as a pre-screen to trigger administration of the QPD.

Using custom programming, the study dataset is then used to generate Bayesian evidence probabilities that are used to perform higher accuracy probabilistic scoring of the dataset.[11] The probabilistic results are used to compute a probabilistic severity score that is mapped back to the conventional PHQ-9 severity score, highlighting the error associated with conventional scoring. Finally simulations are run to map the accuracy of probabilistic scoring to conventional scoring based on the average number of questions administered. This highlights how dynamic administration of the PHQ-9 based on probabilistic scoring, can reduce the time required to administer the PHQ-9 to fidelity, while still enabling additional domains of information to be gathered in an equivalent amount of time.

Results

Looking at the underlying population, the frequency of positive results are summarized in Table 1 for the QPD, PHQ-9 and PHQ-2 at several cut scores. These rates may appear high but are not uncommon compared to similar types of practices.

Table 1



N=2495	Count	Percent
QPD (+)	1101	44.1%
PHQ9 > 9	782	31.3%
PHQ2 > 2	763	30.6%
PHQ2 > 1	1241	49.7%
PHQ2 > 0	1625	65.1%

Table 2 summarizes the false positives and false negatives compared to the QPD, of using the PHQ-9 or PHQ-2 at several cut scores as a pre-screen to trigger the QPD. These results also represent the number of false positives and false negatives when using the PHQ-9 or PHQ-2 as a screener for further evaluation by a healthcare professional, which is currently standard practice in most clinics.

Table 2

False Results	**PHQ(+), QPD(-)**		**QPD(+), PHQ(-)**	
Created	Count	Percent	Count	Percent
PHQ9 > 9	76	3.0%	395	15.8%
PHQ2 > 2	158	6.3%	496	19.9%
PHQ2 > 1	378	15.2%	238	9.6%
PHQ2 > 0	624	25.0%	100	4.0%

Table 3 summarizes the time and cost associated with using the PHQ-9 or PHQ-2 as a pre-screen for further evaluation by a healthcare professional or alternatively automatically triggering an in-depth assessment such as the QPD. The time calculation is based on 1 minute of staff time to administer the PHQ-9 or PHQ-2, score and data enter the result, and 10 minutes of health professional time to do the evaluation. Labor cost is calculated based on staff burdened at $40 per hour and healthcare professionals at $60 per hour. Materials costs are based on $0.15 per PHQ-9 administration on paper, $0.38 service fee per electronic administration of the PHQ-9 and QPD when triggered, and $1.50 license fee per administration of the QPD. These estimates are intentionally low and would be different across various settings and clinical workflows.

Table 3

Administrations Total			**PHQ + BHT interview**			**Electronic**		
Cut Score	QPD tot	PHQ tot	Staff (hrs)	BHT (hrs)	Cost	Staff (hrs)	BHT (hrs)	Cost
PHQ9 > 9	782	2495	41.6	130.3	$9,858	0	0	$2,121
PHQ2 > 2	763	2495	41.6	127.2	$9,668	0	0	$2,093
PHQ2 > 1	1241	2495	41.6	206.8	$14,448	0	0	$2,810
PHQ2 > 0	1625	2495	41.6	270.8	$18,288	0	0	$3,386

Table 4 breaks out the QPD information 1) in the underlying population, 2) when the PHQ-9 is triggered and 3) when the PHQ-9 pre-screen is false negative. These results highlight the information that is missed based on how a practice implements.

Table 4

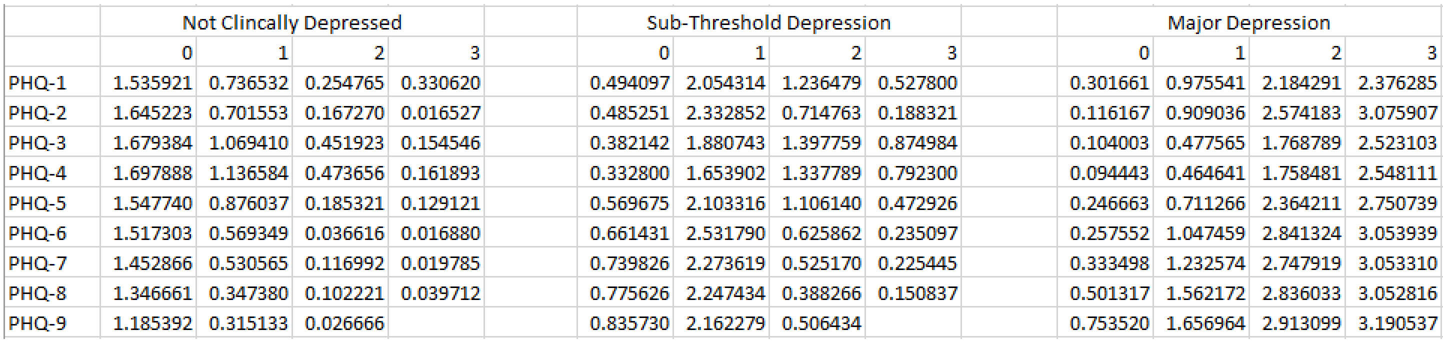
N=2495	Population (n=1101)			PHQ9 > 9 (n=782)			QPD(+), PHQ9(-) (n=395)	
QPD_MajDep	579	23.2%		481	61.5%		98	24.8%
QPD_Dysthym	44	1.8%		25	3.2%		19	4.8%
QPD_Dep_NOS	223	8.9%		142	18.2%		81	20.5%
QPD_Bipolar	442	17.7%		330	42.2%		112	28.4%
QPD_GenAnx	260	10.4%		216	27.6%		44	11.1%
QPD_Panic	161	6.5%		120	15.3%		41	10.4%
QPD_PTSD	739	29.6%		430	55.0%		171	43.3%
QPD_OCD	20	0.8%		18	2.3%		2	0.5%
QPD_Anx_NOS	299	12.0%		112	14.3%		187	47.3%
QPD_Anx2Dep	353	14.1%		290	37.1%		63	15.9%
QPD_Bulimia	29	1.2%		20	2.6%		9	2.3%
QPD_SubAbuse	99	4.0%		60	7.7%		32	8.1%
QPD_Somatiz	220	8.8%		177	22.6%		43	10.9%
QPD_SuicIdea	176	7.1%		151	19.3%		22	5.6%
QPD_SuicRisk	25	1.0%		18	2.3%		7	1.8%

Further, using the study dataset as probabilistic evidence, item response theory (IRT)[12] states there is a functional relationship between each question item and the PHQ-9 result region. In Table 5 we compute this functional relationship using Bayes theorem.

$$P(E|C) = (P(E|C) / P(E)) * P(C)$$

P(E) is the percentage of cases endorsing a specific question item answer and P(E|C) is the percentage of cases endorsing a specific question item answer with a result in a specific sub-region of the PHQ-9 depression latent trait. Table 5 lists P(E|C) / P(E) as a lookup coefficient that can be applied when a specific answer is endorsed, performing probabilistic scoring by updating the a priori probability of the respondent being in each sub-region of the latent trait after each question is answered.

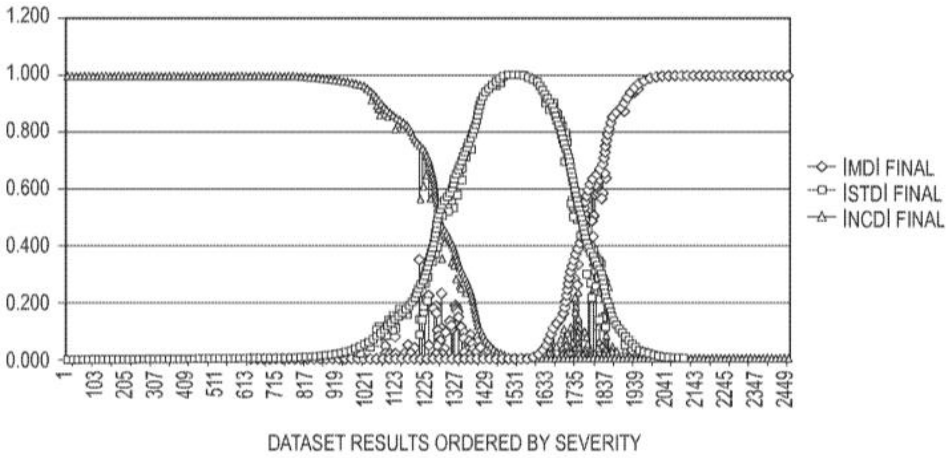
Table 5



	Not Clincally Depressed					Sub-Threshold Depression					Major Depression			
	0	1	2	3		0	1	2	3		0	1	2	3
PHQ-1	1.535921	0.736532	0.254765	0.330620		0.494097	2.054314	1.236479	0.527800		0.301661	0.975541	2.184291	2.376285
PHQ-2	1.645223	0.701553	0.167270	0.016527		0.485251	2.332852	0.714763	0.188321		0.116167	0.909036	2.574183	3.075907
PHQ-3	1.679384	1.069410	0.451923	0.154546		0.382142	1.880743	1.397759	0.874984		0.104003	0.477565	1.768789	2.523103
PHQ-4	1.697888	1.136584	0.473656	0.161893		0.332800	1.653902	1.337789	0.792300		0.094443	0.464641	1.758481	2.548111
PHQ-5	1.547740	0.876037	0.185321	0.129121		0.569675	2.103316	1.106140	0.472926		0.246663	0.711266	2.364211	2.750739
PHQ-6	1.517303	0.569349	0.036616	0.016880		0.661431	2.531790	0.625862	0.235097		0.257552	1.047459	2.841324	3.053939
PHQ-7	1.452866	0.530565	0.116992	0.019785		0.739826	2.273619	0.525170	0.225445		0.333498	1.232574	2.747919	3.053310
PHQ-8	1.346661	0.347380	0.102221	0.039712		0.775626	2.247434	0.388266	0.150837		0.501317	1.562172	2.836033	3.052816
PHQ-9	1.185392	0.315133	0.026666			0.835730	2.162279	0.506434			0.753520	1.656964	2.913099	3.190537

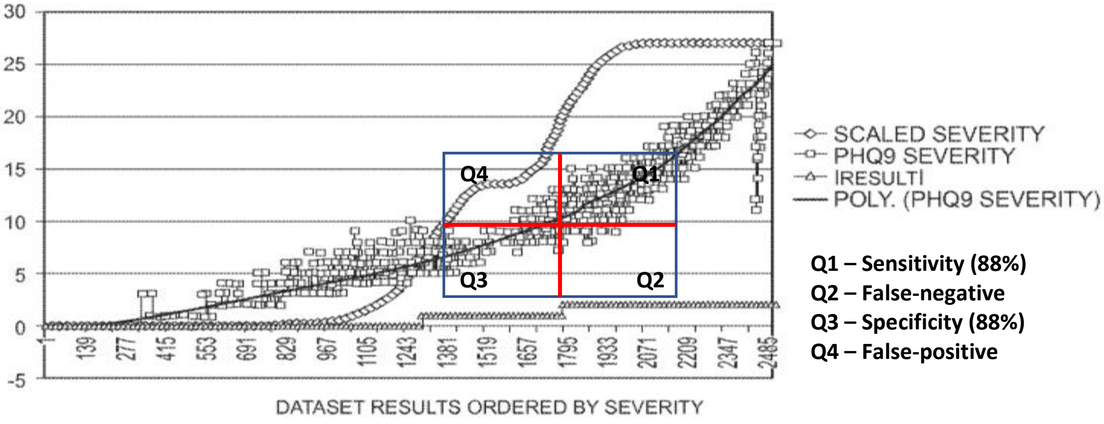
It is important to note that conventional scoring treats the PHQ-9 depression latent trait as a linear scale with a severity score of 0 – 27, where each item is given equal value (3 out of 27) and each answer is given an integer value of equal incremental size (0 – 3). If this was a valid assumption, each column in Table 5 would have the same coefficient value which it does not. Plotting the probabilistic scored results, major depression (MD), sub-threshold depression (STD) and not clinically depressed (NCD), ordered by conventional severity in Figure 1, shows the error introduced by making the assumption of integer, linear values in conventional scoring.

Figure 1



To illustrate this further, using the difference in probabilistic result between latent trait sub-regions, we can construct a probabilistic severity score.[14] In Figure 2 we plot probabilistic scored severity against conventionally scored severity. A polynomial fit curve is applied to the conventional scores and can be used to create a lookup table to convert probabilistic severity scores back to conventional linear scores for reporting into the EMR. To help visualize the associated conventional scoring error, a cross-hair is applied showing this error in the 2nd and 4th quadrants.

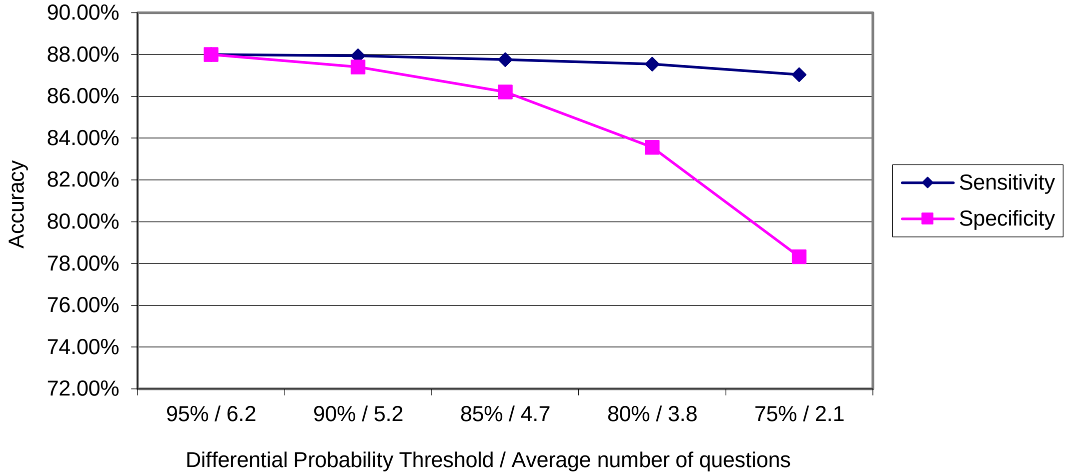
Figure 2



Q1 – Sensitivity (88%)
Q2 – False-negative
Q3 – Specificity (88%)
Q4 – False-positive

In addition to the improved accuracy of probabilistic scoring, each question item is an independent measure of the latent trait. This enables dynamic administration of measures where questions only need to be asked until the result is known within an acceptable level of certainty. Simulating a simple threshold strategy for dynamic administration, Figure 3 summarizes a range of required certainties and the resulting number of questions required. For the PHQ-9, 4.5 questions on average can maintain sensitivity, and within a few percent, the specificity of administering all 9 questions. This could potentially cut the adminstration time required for the PHQ-9 in half on average.

Figure 3



Discussion: Clinical, Operational, and Financial Implications

Making a decision to meet the USPSTF guidelines and simply screen using a PHQ-2 or PHQ-9 on paper or via staff interview, appears to be a low cost, effective screening solution; however, as we outline in this paper, the decision is actually high cost and not always the most effective implementation.

Depth and Breadth

Using Peek's "three world view" focused on clinical, much of the missed information in Table 4 from using the PHQ-9 as a prescreen, comes from anxiety-based issues. Screening is normally integrated in medical settings where anxiety, as well as somatization, are relevant to the medical decision process. Depression can be over 70% comorbid with anxiety, so prescreening with the PHQ-9 detects a large portion of anxiety related issues.[15] However, expanding breadth to include at least a GAD-2 or GAD-7 would help to detect a fuller range of issues. Expanding depth and breadth in general, better supports early detection, meeting guidelines, documentation, clinical decisions and the warm handoff process.

From an operational view, there are tradeoffs around when the in-depth evaluation is performed. A typical scenario is to detect that there are issues, perform a warm handoff where the mental health clinician meets the patient and builds a level of trust before performing the evaluation at the first session. Because of the difficulty getting patients to return for a session, a more effective process being used is to perform the evaluation during the warm handoff. This still has a "one and done" issue so moving the evaluation into the screening process, and simply confirming and then addressing issues during the warm handoff is an even more effective approach.

From a financial view, automated in-depth computer assessment, as shown in Table 3 is far more cost-effective than using mental health clinicians to perform the same work. In addition, screening and computer assessment can be billed separately and the mental health clinicians time that is saved, can be used to generate other sources of revenue. Fully automated screening is considered a profit center in practices where the accounting is setup to track its activity.

Accuracy

From a clinical view, there is clinical screening fatigue from using measures with poor accuracy and shallow information. In a recent study, providers prescreening with the PHQ-2 did not refer a positive result, 95% of the time, primarily because they did not think there was useful information to be found.[8,9] Poor implementation has rendered these screening programs effectively useless.

False positive and false negative results have clinical consequences. Classifying a patient as having an issue when they do not, or alternatively not being aware the patient has an issue, affects clinical decisions and ultimately the health of the patient. Errors in prescreening, highlighted in Table 2, impede the clinical decision process, while errors in the accuracy of scoring measures, highlighted in Figure 2, become part of the patient's medical record. Dynamically administering a prescreen and triggering into an in-depth assessment, effectively eliminates prescreen error in that decisions with consequences are made on full information not the single measure of the PHQ-9.

From an operational view, the consequences of false positive results, whether from prescreening or from measure scoring, are significant. The time spent by staff and BHTs shown in Table 3, following up on the false results requires more personnel with their associated onboarding, training and overhead costs. The unnecessary work required affects patient flow and office operations as well.

From a financial view, as shown in Table 3, the labor costs associated with the unnecessary processing of prescreening false positive results is significant. If you consider false positive results from scoring error at 12% for the PHQ-9, assume every patient is screened annually and apply the same assumptions used in Table 3, the cost per patient for a healthcare system is over $2 per patient on average. The false positive error rate associated with probabilistic scoring is a function of sample size but assuming you are able to cut the error rate in half, it's over a $1 per patient on average that is saved annually.

Automated Processing

From a clinical view, whether you are using prescreens to trigger in-depth measures, or probabilistic scoring to question until adequately certain of the result, the only way to achieve depth and breadth is through dynamic administration. Patients and the screening process can only tolerate a limited amount of questions and duration, before the administration process will start affecting patient flow and office process. AI-based interviewing will be available soon, and will advance dynamic administration even further.

From an operational view, dynamic administration requires screening to be in an electronic format which enables far greater reach on to patient's personal devices and it also gets staff out of the process. Patients do their own data entry, removing personal interview bias and the potential for staff to help answer questions. The issue then becomes how to interface screening results with the EMR where the patient information needs to reside.

What is often misunderstood is that once the data fields and access to the data has been established in the EMR, the patient data entry can be performed by staff, typically form-based, or simply transacted or manually imported electronically as lab data from a 3rd party system. 3rd party systems provide access to advanced screening capabilities such as dynamic administration, and libraries of licensed and public measures available in multiple languages and formats. In addition, 3rd party systems can take scheduling information, apply an age or episode-based screening protocol and manage the screening process, sending notifications, reminders and access links as required. Electronic Medical Records are good at provider-facing interaction, but typically do not support dynamic administration, scoring and reporting and require each practice to build its own library of screening content.

From a financial view, automated processing can significantly reduce labor and its associated cost, both staff labor to administer screenings and mental health labor to perform in-depth evaluation. When set up properly, patients are doing data entry for the practice and resulting information is automatically scored and reported to not just the EMR but optionally to a data warehouse or directly to and from other community resources.

Limitations

For better accuracy, evidence probabilities and item response theory in general, is dependent on large evidence dataset sample size. Often the original validation dataset for measures are not available or when available, the sample size is small. This leads to the need for relatively easy practice-based research, collecting the evidence datasets as part of normal screening. Further, dynamic administration creates the need for obtaining evidence linking one standardized measure to another, for example PHQ-9 items linked to domains of the QPD in this study or linked to domains of suicidality in other work being proposed. This creates more specific triggers into other measures and a more effective dynamic administration.

Discussion Summary

A simple practice-based study combining the PHQ-9 with an in-depth computer assessment has been able to show the high cost and limited effectiveness of current screening implementations. To drive down the cost, these data suggest that practices would benefit from automating their processing and improve the accuracy of their screening. To increase effectiveness, we suggest practices need to expand the depth and breadth of screening, which requires dynamically administering screens. Probabilistic scoring has been introduced as a higher accuracy alternative that enables dynamic administration, making it practical to expand depth and breadth in medical settings.

References

1. Akincigil A, Matthews EB. National Rates and Patterns of Depression Screening in Primary Care: Results From 2012 and 2013. *Psychiatric Services.* 2017;68(7):660-666.
2. James SL, Abate D, Abate KH, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990&#x2013;2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet.* 2018;392(10159):1789-1858.
3. Siu AL, Force atUPST. Screening for Depression in Adults: US Preventive Services Task Force Recommendation Statement. *JAMA.* 2016;315(4):380-387.
4. Gates K, Petterson S, Wingrove P, Miller B, Klink K. You Can't Treat What You Don't Diagnose: An Analysis of the Recognition of Somatic Presentations of Depression and Anxiety in Primary Care. *Families, Systems, & Health.* 2016:No Pagination Specified.
5. Cohen DJ, Balasubramanian BA, Davis M, et al. Understanding Care Integration from the Ground Up: Five Organizing Constructs that Shape Integrated Practices. *The Journal of the American Board of Family Medicine.* 2015;28(Supplement 1):S7-S20.
6. Cohen DJ, Davis MD, Balasubramanian BA, et al. The Dance of Collaboration: Consulting, Coordinating, and Collaborating *Journal of the American Board of Family Medicine* 2015.
7. Kroenke K, Spitzer RL. The PHQ-9: a new depression and diagnostic severity measure. *Psychiatr Ann.* 2002;32.
8. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med.* 2001;16(9):606-613.
9. Fuchs CH, Haradhvala N, Hubley S, et al. Physician actions following a positive PHQ-2: Implications for the implementation of depression screening in family medicine practice. *Families, Systems, & Health.* 2015;33(1):18-27.
10. Peek CJ. Planning care in the clinical, operational, and financial worlds. In: Kessler R, Stafford D, eds. *Collaborative Medicine Case Studies: Evidence in Practice.* New York: Springer; 2008.
11. Knuth KH, Habeck M, Malakar NK, Mubeen AM, Placek B. Bayesian evidence and model selection. *Digital Signal Processing.* 2015;47:50-67.
12. Edelen MO, Reeve BB. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research.* 2007;16(1):5.
13. Shedler, J (2017). Automated mental health assessment for Integrated care: The Quick PsychoDiagnostics Panel meets real-world clinical needs. In RW Feinstein, JV Connely, & MS Feinstein, Eds., *Integrating Behavioral Health and Primary Care.* NY: Oxford University Press; 2017:134-145.
14. Malik AD. Methods and Systems For Assessing Latent Traits Using Probabilistic Scoring. *US Patent No. 8,834,174 B2.* 2014.
15. Olfson M, Fireman B, Weissman MM, Leon AC, Sheehan DV, Kathol RG, Hoven C, Farber L. Mental disorders and disability among patients in a primary care group practice. *Am J Psychiatry.* 1997 Dec; 154(12):1734-40.