ORIGINAL RESEARCH

Screening for Depression in Mobile Devices Using Patient Health Questionnaire-9 (PHQ-9) Data: A Diagnostic Meta-Analysis via Machine Learning Methods

Sunhae Kim ⓘD
Kounseok Lee

Department of Psychiatry, Hanyang University Medical Center, Seoul, Korea

Purpose: Depression is a symptom commonly encountered in primary care; however, it is often not detected by doctors. Recently, disease diagnosis and treatment approaches have been attempted using smart devices. In this study, instrumental effectiveness was confirmed with the diagnostic meta-analysis of studies that demonstrated the diagnostic effectiveness of PHQ-9 for depression using mobile devices.

Patients and Methods: We found all published and unpublished studies through EMBASE, MEDLINE, MEDLINE In-Process, and PsychINFO up to March 26, 2021. We performed a meta-analysis by including 1099 subjects in four studies. We performed a diagnostic meta-analysis according to the PHQ-9 cut-off score and machine learning algorithm techniques. Quality assessment was conducted using the QUADAS-2 tool. Data on the sensitivity and specificity of the studies included in the meta-analysis were extracted in a standardized format. Bivariate and summary receiver operating characteristic (SROC) curve were constructed using the metandi, midas, metabias, and metareg functions of the Stata algorithm meta-analysis words.

Results: Using four studies out of the 5476 papers searched, a diagnostic meta-analysis of the PHQ-9 scores of 1099 people diagnosed with depression was performed. The pooled sensitivity and specificity were 0.797 (95% CI = 0.642–0.895) and 0.85 (95% CI = 0.780–0.900), respectively. The diagnostic odds ratio was 22.16 (95% CI = 7.273–67.499). Overall, a good balance was maintained, and no heterogeneity or publication bias was presented.

Conclusion: Through various machine learning algorithm techniques, it was possible to confirm that PHQ-9 depression screening in mobiles is an effective diagnostic tool when integrated into a diagnostic meta-analysis.

Keywords: diagnostic meta-analysis, depression, Patient Health Questionnaire-9, machine learning, mobile, diagnosis

Introduction

The use of wearable devices via smartphones and tablet personal computers (PCs) has becomes an everyday occurrence. It is estimated that more than 5 billion people have a mobile device, of which more than 60% (77% in developed countries) use a smart device.[1]

Smartphones and wearables help prevent disease symptoms and can provide long-term disease management through passive and short-term sensing in daily life. The devices do this by capturing the physical, mental, and social aspects of

Correspondence: Kounseok Lee
Tel +82 2 2290 8423
Fax +82 2 2298 2055
Email dual@hanyang.ac.kr

human behavior that constitute well-being.[2,3] The accessibility of mobile apps makes it easy to report routine mental health ratings, such as for depression or other mood states, into ecological instantaneous assessment tools.[4–9]

Existing studies have demonstrated that smartphones are a pervasive computing platform that provide a tremendous opportunity to automatically detect depression using collected sensory data. Further, they can be used for effective depression screenings.[10–13] For example, Kolenik and Gams have reported that intelligent cognitive assistant (ICA) technology is used in various fields to imitate human behavior expressed through language models. This technology can be individually tailored to natural language, which has a huge impact on digital mental health services. In particular, ICA can effectively support stress, anxiety, and depression (SAD) by analyzing people's emotional and cognitive phenomena.[14]

Depression is the most frequent in psychiatric disorder studied via mobile devices. It affects more than 300 million people worldwide.[15] Approximately 10–20% of primary care visits are associated with depression, making it the second most common chronic condition observed by primary care physicians.[16] However, primary care physicians identify only 50% of depression cases;[16] therefore, it can remain undiagnosed, leaving many in need of treatment and putting intervention out of reach. Depression is treated with evidence-based therapeutic approaches; however, only 7% of low-income countries and 28% of high-income countries provide interventions for depressed patients.[17] In addition, the early diagnosis of depression with early intervention and treatment is associated with a better prognosis.[18] Thus, screening for the symptoms of depression is an important issue. For example, persuasive technology (PT) proactively intervenes to alleviate stress, anxiety, and depression (SAD), which are critical issues in mental health well-being. These technologies are economical and capable of being used over remote distances; however, their usefulness is still limited.[19] Therefore, preliminary diagnosis appears as an important issue along with prior intervention.

Previous studies on depression have shown correlations between various smartphone use attributes and depressive behaviors.[2,12,20–26] In addition, the development of wristbands and smartphone-embedded sensors over the past

decade has provided an opportunity to objectively measure numerous characteristic symptoms of depression and facilitate the passive monitoring of behavioral indicators of low mood.[27]

Individual behavioral characteristics that discriminate depressive symptoms include physical activity (eg, walking, running, sleeping[28]), behavioral changes (eg, smartphone conversation patterns such as language fluency and intonation[29]), circadian activity, and social interaction. These can be detected through a smartphone sensor in association with lassitude, anesthesia, and psychomotor retardation. Depressed people, for example, appear to make fewer phone calls and search the Internet less frequently on their mobile phones.[16] A mobile GPS can help assess the severity of depression by movement.[28,30–32]

These technologies serve as an investigation method that discriminates depressive symptoms based on the characteristics of individual behaviors derived from the monitoring sensor.[33–38] They use a smartphone sensor to distinguish depressed people from non-depressed people.[12,30–32,36,39] However, studies to screen for and diagnose depression based on differences in levels of depression are lacking.[39,40] Therefore, a study targeting different patient groups is necessary for the diagnosis of depression.

A diagnosis of depression is traditionally performed as a paper-type self-report tool. Diagnosis through a smartphone uses a validated self-report screening tool along with passive monitoring.[12,41–43] This self-report screening tool is simple, economical, and familiar to people. The Patient Health Questionnaire (PHQ) used in the self-reported depression diagnosis is a depression self-report tool. It is a new tool for the criteria-based diagnosis of depression and other psychiatric disorders commonly encountered in primary care. It is half the length of many other measures of depression, with similar sensitivity and specificity. It consists of nine real-world criteria that form the basis for the diagnosis of DSM-IV Depressive Disorder.[44] It is a validated measure of depression often used as a screening tool in clinical settings.[45–49] Numerous studies have demonstrated the usefulness of PHQ-9 in influencing clinical decision-making.[50] Moreover, a meta-analysis has concluded that shows its superior diagnostic properties when compared with longer screening tools.[51]

In general, the results of self-report questionnaires, including PHQ-9, are often used as a discriminate tool

for depressive symptoms.[52] In particular, PHQ-9 has a diagnostic basis for DSM-IV and can be used in various fields because it consists of short questions. In addition, the level of depression indicated by the PHQ-9 score shows a high correlation with the features of depression that can be detected by a smartphone.[12,42,53–55]

In addition to the merits of PHQ-9 as a discriminate tool for depression, it has high sensitivity and specificity. In a primary care study PHQ-9 data, the algorithm sensitivity and specificity was 73% and 98%,[56] respectively. In the validation study for the summed-item method, a PHQ-9 score of ≥10 was 88% for both sensitivity and specificity for major depressive disorder.[44] In recent studies, diagnostic measures [eg, sensitivity, specificity, area under the receiver operating characteristic curve (AUC)] are implemented in machine learning by extending the existing statistical approach for the selection criteria and prediction of PHQ-9.[13,47,54,57–59] Machine learning is used in psychiatry to increase the accuracy of diagnosis and prognosis and make treatment and prevention decisions.[60,61] It is particularly useful for predicting human behavior, including high-risk behaviors, and it is effective for discriminating psychopathology.[62,63] Machine learning studies are conducted using measures to discriminate psychopathology (eg, prediction suicide ideation, suicide attempt and behaviors, malingering, personality detecting).[61,63–69] In addition, studies using PHQ-9 are emerging. These have strengths in the identification of depression by using machine learning techniques.[53,54,58,59,70,71]

In this study, studies were analyzed using machine learning techniques on PHQ-9 depression screening data collected through mobile devices according to the latest depression screening trends. In addition, we intend to contribute to predicting depression through mobile devices in the future by examining the predictive diagnostic power of PHQ-9 on mobile devices through diagnostic meta-analysis.

Materials and Methods
Data Sources and Searches
We searched EMBASE, MEDLINE, MEDLINE In-Process, and PsychINFO between 1964 and March 26, 2021. "Depression," "depressive disorder," "mood disorder," and "sensing," "sensor," "measuring," and "diagnosis" related to sensing were used as keywords. Combined expressions were searched. Below is the

mutually agreed upon some search query, and additional search query is presented in the supplementary (Supplementary Materials).

"depression" AND "smartphone" OR "depression" AND "wearable" OR "depression" AND "mobile" OR "depression" AND "smartphone" AND "sensing" OR "depression" AND "smartphone" AND "sensor" OR "depression" AND "smartphone" AND "measuring" OR "depression" AND "smartphone" AND "diagnosis".

We included review articles, posters, all kinds of unpublished studies, and studies without language restrictions. This study was prepared according to the preferred reporting items for systematic reviews and meta-analyses (PRISMA) guidelines and completed 27-point checklist PRISMA.[72]

Study Selection
After removing 1340 duplicate papers from the 5476 papers searched, two reviewers independently first selected the papers by title and abstract. All authors applied the eligibility criteria of quality assessment using the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tool and PRISMA checklist. They screened and reviewed the entire text of the papers and created a final list of papers containing eligible data through consensus (Figure 1). We removed all reports of mental disorders (such as bipolar disorder, and schizophrenia), except for depression because of other psychiatry mental discriminant categories. We excluded studies and interventions that discriminated mental health through sensing. In the case of interventions through sensing, there were studies that verified the therapeutic effects of interventions; therefore, an integrated analysis through meta-analysis was not possible, so this study was excluded. Only the depressed group as selected by the PHQ-9 were included. Those selected by the PHQ-2 and the PHQ-8, which are short-cut scales of the PHQ, were excluded. In addition, methodologies that selected or measured the PHQ-9 by traditional statistical methods other than machine learning were excluded.

Statistical Analysis
Data Synthesis
The diagnostic accuracy was extracted in a standardized format with all possible participant characteristics, scores on PHQ-9, and data on sensitivity and specificity. Where appropriate, the cell contents of the 2×2 table
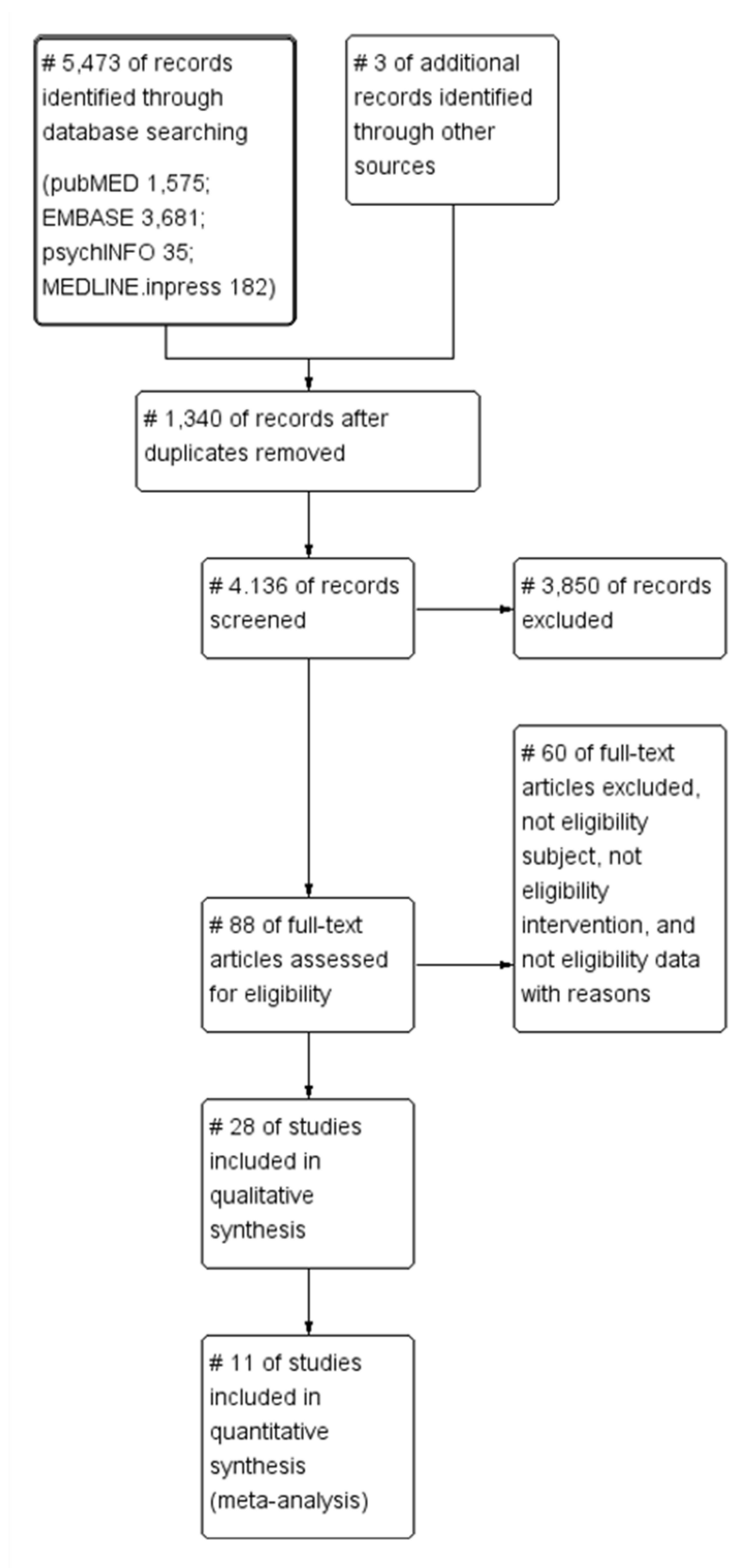
Figure 1 Flow diagram (PRISMA).
Notes: PRISMA figure adapted from Moher D, Liberati A, Altman D, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. Journal of clinical epidemiology. 2009;62(10). Creative Commons.[72]

Table 1 Demographic Contents Include Studies

Study	Age	Sample Size	PHQ-9 Characteristics (Depressed Severity Level)	PHQ-9 Interval Conducted	Data Collect	Machine Learning Algorithm Method
Dogrucu 2020a[54]	Age: at least 18 years or older	N= 335	Cutoff 10 = moderate depression	2 weeks	Smartphone, Social media data	Random forest
Dogrucu 2020b[54]			Cutoff 15 = moderately severe depression	2 weeks	Smartphone, Social media data	Random forest
Dogrucu 2020c[54]			Cutoff 20 = severe depression	2 weeks	Smartphone, Social media data	Random forest
Masud 2020a[58]	Age: above 18 years of age [mean=24y/SD±5]	N= 33	$10 \leqq$ PHQ-9 <15: moderate depression	Every week	Mobile sensor data (11 weeks)	Support vector machine (SVM)
Masud 2020b[58]			$10 \leqq$ PHQ-9 <15: moderate depression	Every week	Mobile sensor data (11 weeks)	K-nearest neighbor (KNN)
Masud 2020c[58]			$10 \leqq$ PHQ-9 <15: moderate depression	Every week	Mobile sensor data (11 weeks)	Artificial neural network (ANN)
Masud 2020d[58]			PHQ-9 $\geqq$ 15: severe depression	Every week	Mobile sensor data (11 weeks)	Support vector machine (SVM)
Masud 2020e[58]			PHQ-9 $\geqq$ 15: severe depression	Every week	Mobile sensor data (11 weeks)	K-nearest neighbor (KNN)
Masud 2020f[58]			PHQ-9 $\geqq$ 15: severe depression	Every week	Mobile sensor data (11 weeks)	Artificial neural network (ANN)
Piette 2013[71]	Age: average 52.2 years [SD=12.5]	N= 208	PHQ-9 $\geqq$ 10: moderate/severe	2 weeks (weekly, biweekly, monthly)	IVR (interactive voice response)	10-fold cross validation
McIntyre 2021[59]	Age: 18–65 [mean=46y ±12.7]	N= 523	PHQ-9 $\geqq$ 5: depressed	14 days	Mobile phone on Android platform	10-fold cross validation

Table 2 QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies-2)

Study	Patient Selection: Consecutive or Random Sample of Enrolled?	Patient Selection: Avoid Case-Control Design	Patient Selection: Avoided Inappropriate Exclusions?	Patient Selection: Overall Risk of Bias	Patient Election: Concerns Regarding Applicability	Index Test: Index Test Results Interpreted Without Knowledge of the Results of the Reference Standard?	Index Test: If Threshold Pre-Specified	Index Test: Overall Risk of Bias
Dogrucu 2020a[54]	◎	○	◎	◎	◎	○	○	▽
Dogrucu 2020b[54]	◎	○	◎	◎	◎	○	○	▽
Dogrucu 2020c[54]	◎	○	◎	◎	◎	○	○	▽
Masud 2020a[58]	◎	◎	◎	◎	◎	○	○	▽
Masud 2020b[58]	◎	◎	◎	◎	◎	○	○	▽
Masud 2020c[58]	◎	◎	◎	◎	◎	○	○	▽
Masud 2020d[58]	◎	◎	◎	◎	◎	○	○	▽
Masud 2020e[58]	◎	◎	◎	◎	◎	○	○	▽
Masud 2020f[58]	◎	◎	◎	◎	◎	○	○	▽
Piette 2013[71]	◎	◎	◎	◎	◎	○	○	▽
McIntyre 2021[59]	◎	◎	○	◎	◎	○	○	▽

Abbreviations: ○, Yes; ●, No; ▽, Low; ◎, Unclear; ▲, High.

were used by analyzing the receiver operating characteristic (ROC) curve calculated from the provided data and plotted.[73]

Meta-Analysis

We performed a bivariate meta-analysis to obtain pooled estimates of specificity and sensitivity and 95% confidence intervals (CIs) and generate 95% confidence ellipses within the ROC curve space.[74] A summary receiver operating characteristic (SROC) curve was constructed using a quantitative model.[75]

The heterogeneity was evaluated using I^2 (I^2; the proportion of true variance), and meta-regression was performed to determine the heterogeneity. In addition, publication bias was assess to determine any bias in which a study may or may not be published according to the characteristics and results of individual studies. The probability of having the disease in question was estimated based on the diagnostic test results through Fagan's nomogram.[76] Analyses were performed using STATA 17.0 (Texas, USA) using metandi, midas, meta-bias, and metareg of Stata algorithm meta-analysis words. "metandi" performs meta-analysis of diagnostic accuracy and it takes as input four variables: tp (true positives), fp (false positives), fn (false negatives), tn (true negatives) within each study.[77] "midas" is a comprehensive program of statistical an graphical routines for undertaking meta-analysis of diagnostic test performance in Stata.[78] And "metareg" performs

Table 3 Meta-Analysis of Diagnostic Accuracy

Variable		Coef	Std Err	z	P	95% Conf Interval
Bivariate	Corr (logits)	0.7630635	0.1940425			0.0928132–0.9574147
HSROC	Beta	−0.5151134	−0.2855053	−1.80	0.071	−1.074693–0.0444667
Summary pt	Sensitivity	0.7965256	0.644144			0.6423612–0.895089
	Specificity	0.8498525	0.0303055			0.7803928–0.9001536
	DOR	22.15723	12.59309			7.273342–67.49893
	LR+	5.304954	1.363503			3.205546–8.779327
	LR-	0.2394232	0.0810192			0.123346–0.4647368
	1/LR-	4.176705	1.413369			2.151756–8.107272

Notes: Log likelihood = −58.743526; Number of studies = 11; Covariance between estimates of E(logitSe) and E(logitSp) = 0.0543045.

Index Test: Concerns Regarding Applicability	Reference Test: Reference Test Correctly Classifies Target Condition	Reference Test: Reference Standard Results Interpreted Blind to Index Test	Reference Test: Overall Risk of Bias	Reference Test: Concerns Regarding Applicability	Flow/Timing: Appropriate Interval Between Index Test and Reference Standard	Flow/Timing: All Participants Receive Same Reference Test	Flow/Timing: All Participants Included in Analysis	Flow/Timing: Overall Risk of Bias	Flow/Timing: Concerns Regarding Applicability
▽	◎	●	◎	◎	○	○	○	▽	▽
▽	◎	●	◎	◎	○	○	○	▽	▽
▽	◎	●	◎	◎	○	○	○	▽	▽
▽	◎	●	◎	◎	◎	○	○	▽	▽
▽	◎	●	◎	◎	◎	○	○	▽	▽
▽	◎	●	◎	◎	◎	○	○	▽	▽
▽	◎	●	◎	◎	◎	○	○	▽	▽
▽	◎	●	◎	◎	◎	○	○	▽	▽
▽	◎	●	◎	◎	◎	○	○	▽	▽
▽	◎	●	◎	◎	○	○	○	▽	▽
▽	◎	●	◎	◎	○	○	○	▽	▽

random-effects meta-regression using aggregate-level data.[79]

Quality Assessment

We conducted a quality assessment using the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tool.[80] This integrates an assessment of bias risk across four key areas: patient selection, index test, reference standard, and flow and timing of assessments. Two authors independently assessed the risk of bias based on consensus criteria.

Results
General Characteristics and Quality Analysis

We performed a meta-analysis of 1099 patients in four studies (age above 18 years old, moderate, and severe depression, data collection through mobile devices, Data analysis through machine learning algorithms). We performed a diagnostic meta-analysis according to the PHQ-9 cutoff score and machine learning algorithm techniques. The machine learning algorithm techniques were random forest, support vector machine (SVM), k-nearest neighbor (KNN), artificial neural network (ANN), and 10-fold cross-validation. Random forest (RF) is an reliable classifier that uses predictions derived from ensembles of decision trees,[81] which successfully select and rank variables with the greatest ability to discriminate between the target classes.[82] The Support Vector Machine is a discriminant classifier that can be defined as a separating hyperplane. It is the generalization of maximal margin classifier that comes with the definition of hyperplane.[81] The K-Nearest Neighbors (kNN) algorithm is used for classification and regression. It performs great in pattern recognition and predictive analysis.[83] The artificial neural network (ANN) is a machine learning method evolved from the idea of simulating the human brain.[84] It is excellent fault tolerance and is fast and highly scalable with parallel processing.[85] Cross-validation is widely used to estimate the prediction error, and also been used for model selection.[86] These machine learning algorithm techniques analyzed PHQ-9 data collected by mobile devices, and this study performed diagnostic meta-analysis by integrating these studies.

The depression diagnosis cutoff score used with the PHQ-9 was 10 points for most studies, 15 points for severe cases, and 5 points or more as a depression diagnosis in one study (Table 1). Therefore, this study performed meta-regression and publication bias according to the level of depression. In addition, a quality analysis of the papers that were analyzed in a meta-

analysis according to QUADAS-2 was performed (Table 2).

Diagnostic Meta Results of PHQ-9 Using Machine Learning

Among the summarized estimates, the pooled sensitivity was 0.797 (95% CI = 0.642–0.895), the pooled specificity was 0.850 (95% CI = 0.780–0.900), and the diagnostic odds ratio (OR) was 22.16 (95% CI = 7.273–67.499). The p-value of HOROC beta was not significant (0.071). If the HOROC beta value is a parameter representing the shape of the SROC curve and is statistically significant, heterogeneity is suspected. Therefore, this study did not show heterogeneity. However, the Corr (logits) correlation coefficient of the bivariate model was 0.763. To confirm heterogeneity between studies, the heterogeneity value was analyzed using a meta-regression analysis (Table 3). Heterogeneity refers to the degree of dispersion of effect sizes from each individual study and the degree of

inconsistency in effect sizes across studies. This heterogeneity is important in meta-analysis to increase the relevance of conclusions drawn from subject studies and to improve scientific understanding of the evidence as a -whole.[87,88] As such, meta-regression analysis performs whether between-study heterogeneity can be explained by one or more moderators.[89,90] Therefore, in this study, the cause of the heterogeneity was further tested through meta-regression analysis.

By analyzing the diagnostic meta result value for the diagnostic value of PHQ-9 using machine learning with a forest plot, the inter- and intra-study variation were confirmed. The intra-study variation was relatively large compared with that of previous studies (sensitivity = 0.50 [95% CI; 0.01–0.99], specificity = 0.83 [95% CI; 0.63–0.95]). McIntyre's study showed a large inter-study variation in both sensitivity and specificity compared with other studies; however, it showed a good balance overall (Figure 2).
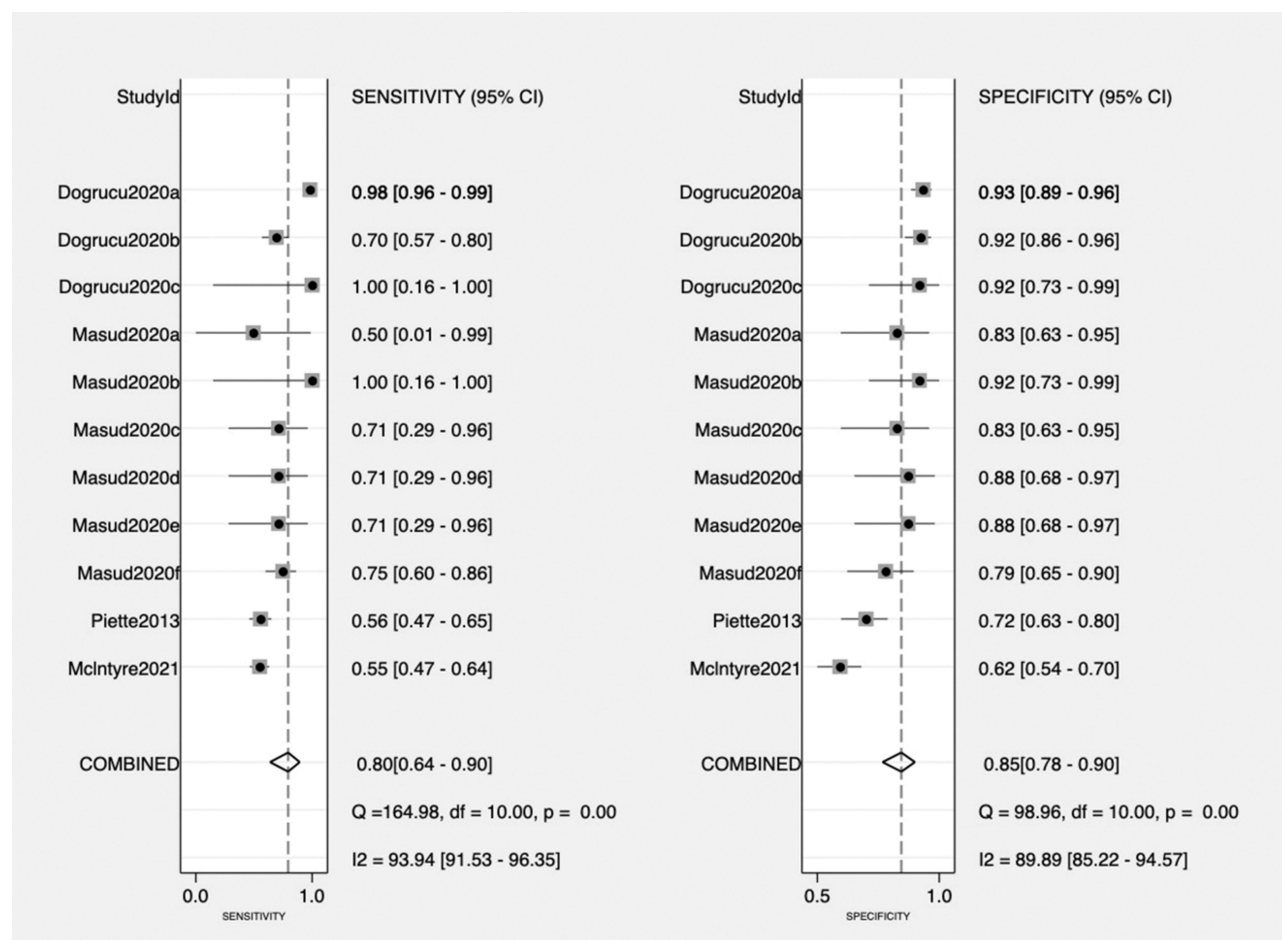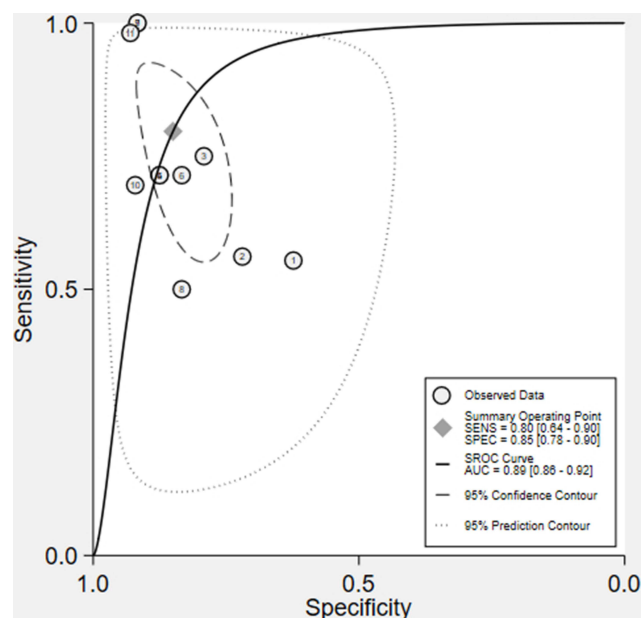

Figure 2 Forest plot.

Figure 3 SROC curve.

Heterogeneity of Diagnostics Meta-Analysis

Although a large amount of heterogeneity is not suspected from the summary estimate and the SROC curve (Figure 3), the summary estimate Corr (logits) correlation coefficient indicates a negative value exceeding 0; therefore, so the PHQ-9 cutoff value was analyzed through meta-regression analysis (Table 4, Figure 4). As a result of the meta-regression analysis, the p-value of the PHQ-9 severe group was 0.95, which was not significant. This confirmed that the PHQ-9 score was not the cause of the heterogeneity.

Publication Errors

The publication errors of the studies used were distributed non-biased asymmetrically based on the regression line.

The p-value was 0.50, which did not indicate publication errors (Figure 5).

Diagnosed with Depression

Fagan's nomogram is a graphical tool that can measure the probability of having a disease based on the results of a diagnostic test using Bayes' theorem that describes the probability of an event.[76,91] Prior probability is entered to calculate the posterior probability (probability of contracting a specific disease). In the case of depression screening, this appears between 5% and 10% in primary care.[92,93] In the meta-analysis study of Mitchell et al that screened for depression using PHQ-2 and PHQ-9, the prevalence in primary care was 11.3% (95% CI 10.92–11.68%).[94] Therefore, assuming that the pre-prevalence value selected with PHQ-9 is 11%, the posterior probability of being diagnosed with depression is 40% if LR_positive, according to this Fagan's nomogram diagnostic test. If the diagnosis is LR_negative, the probability of being diagnosed with depression appears to be 3% (Figure 6).

Discussion

This study is the first diagnostic meta-analysis to reveal the efficacy of depression screening when used with a computer-based mobile device and PHQ-9. In addition, it is the first study to determine through meta-analysis whether machine learning algorithm methods have a diagnostic strength when PHQ-9 is used over traditional statistical methods via machine learning techniques that have predictive power.

The PHQ-9 has diagnostic properties for major depressive disorder.[47,51,94–99] It is a short period and gold standard screening tool for depression based on the DSM-IV diagnostic criteria.[44,56,100,101] In addition, the PHQ-9 is the computer

Table 4 Meta-Regression

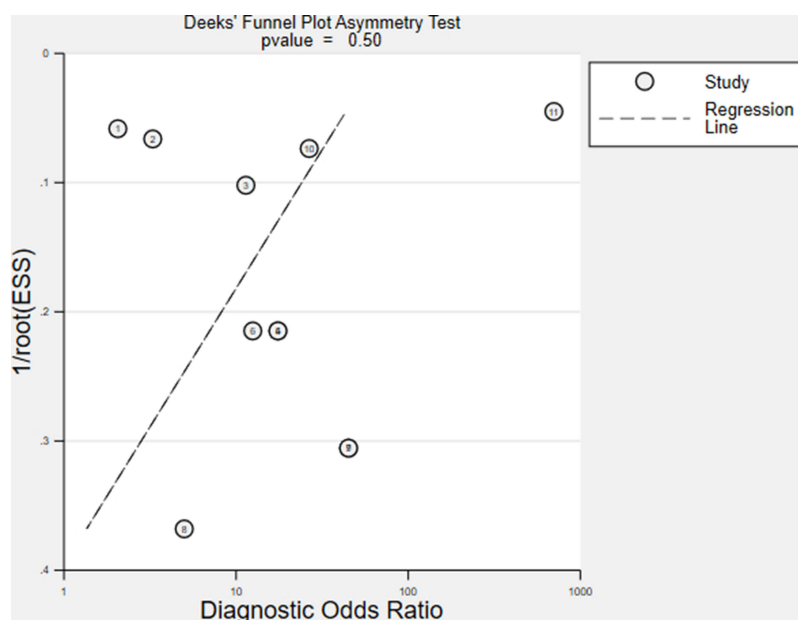| Sensitivity and Specificity						
Parameter	Category	N studies	Sensitivity	P1	Specificity	P2
Phq9severe	Yes	5	0.78 [0.56–1.00]	0.62	0.84 [0.74–0.93]	0.03
	No	6	0.81 [0.66–0.96]		0.86 [0.78–0.93]	
Joint Model						
Parameter	Category	LRTChi2	P value	I^2	I^2lo	I^2hi
Phq9severe	Yes	0.10	0.95	0	0	100
	No					

Figure 4 Meta-regression.

version, and the diagnostic reliability and effectiveness are identical to offline methods[102,103] and with smartphones.[9,53,104]

Similar to the depression diagnostic values of PHQ-9 through the existing traditional technique in this paper,[105] when PHQ-9 data from mobile devices are analyzed with a machine learning algorithm, good results are reported with 80% sensitivity and 85% specificity. This indicates good diagnostic properties.

The depression-diagnostic properties of PHQ-9 are similar when measurements are performed through mobile devices and machine learning techniques are applied. The PHQ-9 cutoff score is optimally ≥10. When meta-regression was performed by dividing according to the level of PHQ-9, five studies assessing severe depression (PHQ-9 ≥15) showed a sensitivity of 78% and specificity of 84%. Six studies of non-severe depression that did not exceed 15 points showed a sensitivity and specificity similar to the overall sensitivity and specificity, of 81% and 86% respectively. Therefore, a PHQ-9 cutoff of ≥10 has more discriminating power to diagnose depression. This is in line with previous studies.[51,106] Taken together, this study verified the diagnostic discriminatory power of depression according to the PHQ-9 of ≥10 cutoff.

Previous studies on diagnostic meta-analysis in various settings using PHQ have shown high sensitivity and specificity. In the case of PHQ-9, the sensitivity was 0.80–0.82 (95% CI 0.71–0.89) and the specificity was 0.84–0.92 (95% CI 0.80–0.95).[51,94] As such, this study confirmed that PHQ-9 shows similar high sensitivity and specificity even with machine learning statistical techniques in a mobile environment.

Diagnosing depression by collecting depressive data from individuals ≥18 years of age in a mobile environment is more meaningful than diagnosing depression in primary care or clinical care settings. Screening for major depression in the mobile environment has been carried out in the general population (eg, university students, college students, the general community[12,30,42,52,53,107,108]), clinical field, and primary care field.[104,109–111] Furthermore, depression diagnoses are performed for those examined in previous studies[32,112,113] and participants recruited through an app.[114,115] The strength of depression screening in a mobile environment is that the group is not limited to one environment; therefore, more data can be obtained. In particular, self-report assessments in a mobile environment support patients to overcome the locational limitations of traditionally managed assessments (ie, paper-and-pencil-based).[6] Moreover, they can collect passive data to reveal the any correlation strengths. An additional potential benefit of depression screening in a mobile environment is that the app can be developed and used based on the storage of mobile devices, portable accessibility, and time-sensitive local and push notifications.[4]
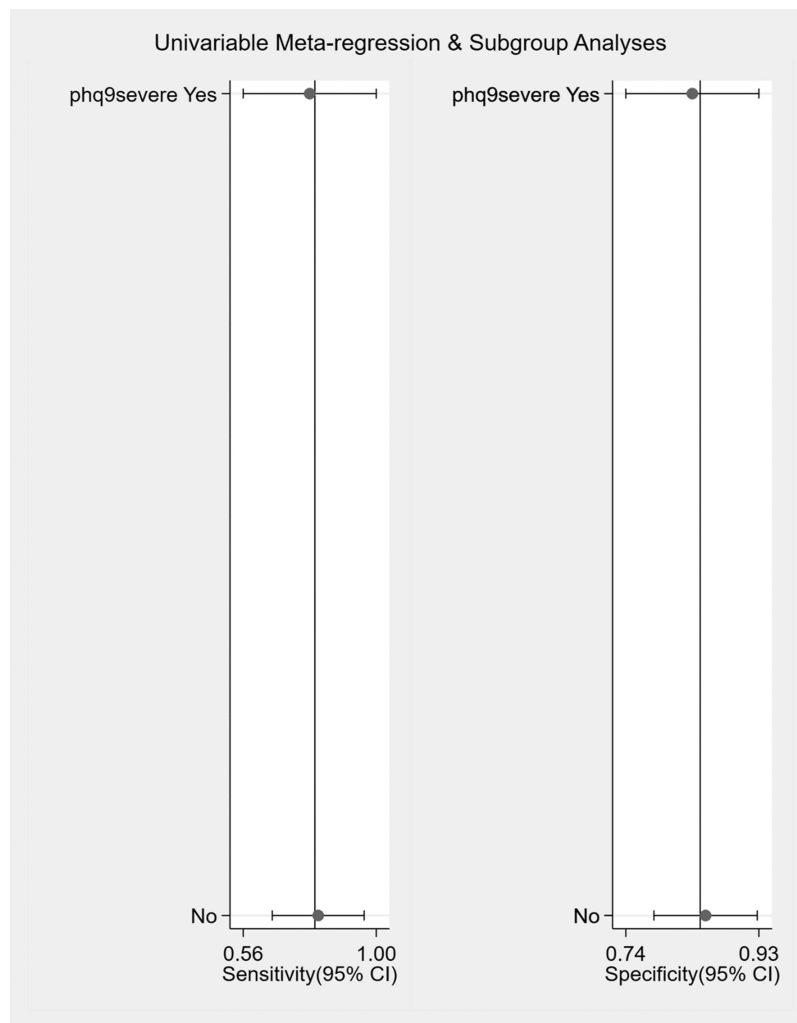
Figure 5 Publication bias.

The strength of this paper is that it shows the diagnostic meta-analysis of machine learning methods with the usefulness of the PHQ-9 for depression screening through mobile devices. Each machine learning algorithm showed little heterogeneity and good diagnostic usefulness. Machine learning selects algorithms according to research and sets up suitable training and testing sets. In this diagnostic meta-analysis, various ML models were analyzed and showed a high fit of AUC = 0.89. For the diagnosis of depression and mood disorders, machine learning with excellent predictive suitability has been introduced.[57,110,113,116,117] Datasets collected in mobile settings are large; machine learning-based predictive models can analyze a large amount of data. This technique is useful for analyzing and conceptualizing multiple predictors.[118,119] In the case of depression, if a diagnosis is delayed by screening tests, the prognosis deteriorates,[120]

and the importance of early detection is raised.[121–124] In addition, as efficiency such as the cost-effectiveness of testing is proven,[125] machine learning and depression screening in the mobile field show potential strengths.

The limitation of this paper is that there are various research results on depression screening conducted in the mobile field, but the results for PHQ-9 are sporadic, and in particular, there are few studies that use machine learning algorithm techniques. In addition, PHQ-9 has excellent diagnostic properties for depression; however, studies on its utility in DSM-5 should be conducted as it is based on DSM-IV. Additionally, the reference standards for diagnosing depression are narrow in the mobile field. Based on the effectiveness of machine learning algorithms for the diagnosis of depression in the mobile field of this study, we expect machine learning studies on depression diagnosis using various sensing data and self-report tests will be collected.
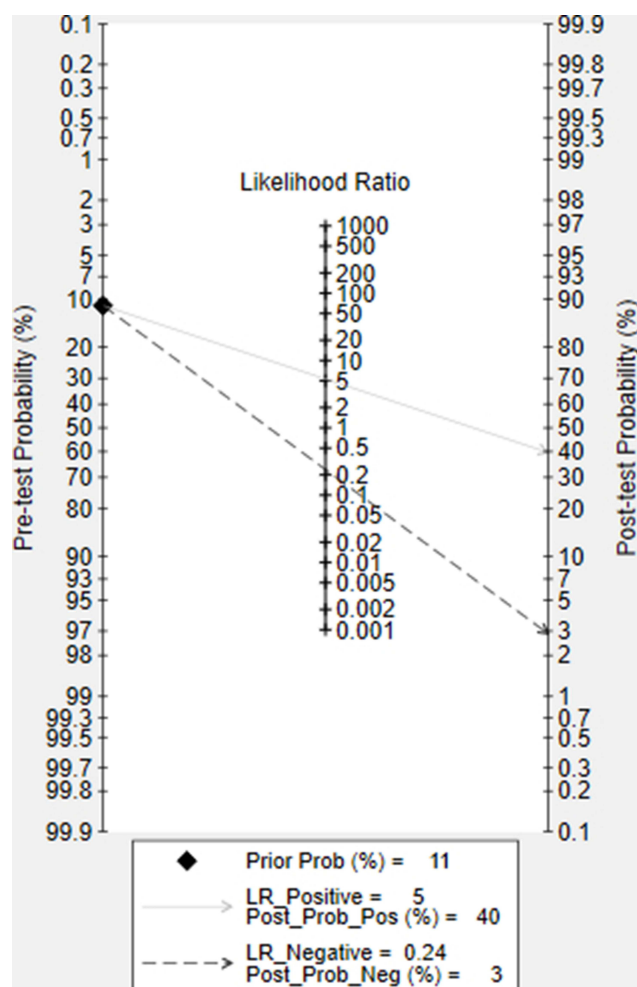
Figure 6 Fagan's nomogram.

Conclusion

In this study, a diagnostic meta-analysis was performed on the selection of depression using machine learning techniques on data related to the PHQ-9, which diagnoses depression in the mobile field. We used mobile data from 1099 subjects with a pooled sensitivity of 80%, specificity of 85%, and an AUC = 0.89 for various machine learning algorithm methods. We found, excellent diagnostic effects by integrating meta-analysis. This study confirms that PHQ-9, which has been proven to be a useful screening test for depression, is an effective diagnostic tool in mobile assessments, as well as in primary care and clinical settings.

Funding

This research was supported by the Technology Innovation Program (20012931) funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea). This funding source had no role in the design of this study and will

not have any role during its execution, analyses, interpretation of the data, or decision to submit results.

Disclosure

The authors declare no conflict of interest.

References

1. Taylor K, Silver L. *Smartphone Ownership is Growing Rapidly Around the World, but Not Always Equally.* Pew Research Center; 2019:5.
2. Burns MN, Begale M, Duffecy J, et al. Harnessing context sensing to develop a mobile intervention for depression. *J Med Internet Res*. 2011;13(3):e55. doi:10.2196/jmir.1838
3. Kasckow J, Zickmund S, Rotondi A, et al. Development of telehealth dialogues for monitoring suicidal patients with schizophrenia: consumer feedback. *Community Ment Health J*. 2014;50 (3):339–342. doi:10.1007/s10597-012-9589-8
4. Donker T, Petrie K, Proudfoot J, Clarke J, Birch M-R, Christensen H. Smartphones for smarter delivery of mental health programs: a systematic review. *J Med Internet Res*. 2013;15(11): e247. doi:10.2196/jmir.2791
5. Hetrick SE, Robinson J, Burge E, et al. Youth codesign of a mobile phone app to facilitate self-monitoring and management of mood symptoms in young people with major depression, suicidal ideation, and self-harm. *JMIR Mental Health*. 2018;5 (1):e9. doi:10.2196/mental.9041
6. Kim J, Lim S, Min YH, et al. Depression screening using daily mental-health ratings from a smartphone application for breast cancer patients. *J Med Internet Res*. 2016;18(8):e216. doi:10.2196/jmir.5598
7. Nahum M, Van Vleet TM, Sohal VS, et al. Immediate mood scaler: tracking symptoms of depression and anxiety using a novel mobile mood scale. *JMIR mHealth uHealth*. 2017;5(4): e6544. doi:10.2196/mhealth.6544
8. Rickard N, Arjmand H-A, Bakker D, Seabrook E. Development of a mobile phone app to support self-monitoring of emotional well-being: a mental health digital innovation. *JMIR Mental Health*. 2016;3(4):e6202. doi:10.2196/mental.6202
9. Torous J, Staples P, Shanahan M, et al. Utilizing a personal smartphone custom app to assess the patient health questionnaire-9 (PHQ-9) depressive symptoms in patients with major depressive disorder. *JMIR Mental Health*. 2015;2(1):e8. doi:10.2196/mental.3889
10. Canzian L, Musolesi M. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. Paper presented at: Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing; 2015.
11. Farhan AA, Lu J, Bi J, Russell A, Wang B, Bamis A. Multi-view bi-clustering to identify smartphone sensing features indicative of depression. Paper presented at: 2016 IEEE first international conference on connected health: applications, systems and engineering technologies (CHASE); 2016.
12. Saeb S, Zhang M, Karr CJ, et al. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *J Med Internet Res*. 2015;17(7):e175. doi:10.2196/jmir.4273
13. Wang R, Chen F, Chen Z, et al. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. Paper presented at: Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing; 2014.

14. Kolenik T, Gams M. Intelligent cognitive assistants for attitude and behavior change support in mental health: state-of-the-art technical review. *Electronics*. 2021;10(11):1250. doi:10.3390/electronics10111250

15. World Health Organization. *Depression and Other Common Mental Disorders: Global Health Estimates*. World Health Organization; 2017.

16. Mitchell AJ, Vaze A, Rao S. Clinical diagnosis of depression in primary care: a meta-analysis. *Lancet*. 2009;374(9690):609–619. doi:10.1016/S0140-6736(09)60879-5

17. Chisholm D, Sweeny K, Sheehan P, et al. Scaling-up treatment of depression and anxiety: a global return on investment analysis. *Lancet Psychiatry*. 2016;3(5):415–424. doi:10.1016/S2215-0366(16)30024-4

18. Kamphuis MH, Stegenga BT, Zuithoff NP, et al. Does recognition of depression in primary care affect outcome? The PREDICT-NL study. *Fam Pract*. 2012;29(1):16–23. doi:10.1093/fampra/cmr049

19. Kolenik T, Gams M. Persuasive technology for mental health: one step closer to (Mental health care) equality? *IEEE Technol Soc Mag*. 2021;40(1):80–86. doi:10.1109/MTS.2021.3056288

20. BinDhim NF, Shaman AM, Trevena L, Basyouni MH, Pont LG, Alhawassi TM. Depression screening via a smartphone app: cross-country user characteristics and feasibility. *J Am Med Inform Assoc*. 2015;22(1):29–34. doi:10.1136/amiajnl-2014-002840

21. Gravenhorst F, Muaremi A, Bardram J, et al. Mobile phones as medical devices in mental disorder treatment: an overview. *Pers Ubiquitous Comput*. 2015;19(2):335–353. doi:10.1007/s00779-014-0829-5

22. Harwood J, Dooley JJ, Scott AJ, Joiner R. Constantly connected–the effects of smart-devices on mental health. *Comput Human Behav*. 2014;34:267–272. doi:10.1016/j.chb.2014.02.006

23. Kumar S, Abowd GD, Abraham WT, et al. Center of excellence for mobile sensor data-to-knowledge (MD2K). *J Am Med Inform Assoc*. 2015;22(6):1137–1142. doi:10.1093/jamia/ocv056

24. Thomée S, Dellve L, Härenstam A, Hagberg M. Perceived connections between information and communication technology use and mental symptoms among young adults-a qualitative study. *BMC Public Health*. 2010;10(1):1–14. doi:10.1186/1471-2458-10-66

25. Thomée S, Härenstam A, Hagberg M. Mobile phone use and stress, sleep disturbances, and symptoms of depression among young adults-a prospective cohort study. *BMC Public Health*. 2011;11(1):1–11. doi:10.1186/1471-2458-11-66

26. Torous J, Friedman R, Keshavan M. Smartphone ownership and interest in mobile applications to monitor symptoms of mental health conditions. *JMIR mHealth uHealth*. 2014;2(1):e2. doi:10.2196/mhealth.2994

27. Torous J, Kiang MV, Lorme J, Onnela J-P. New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. *JMIR Mental Health*. 2016;3(2):e16. doi:10.2196/mental.5165

28. Craft LL, Perna FM. The benefits of exercise for the clinically depressed. *Prim Care Companion J Clin Psychiatry*. 2004;6(3):104. doi:10.4088/PCC.v06n0301

29. George LK, Blazer DG, Hughes DC, Fowler N. Social support and the outcome of major depression. *Br J Psychiatry*. 1989;154(4):478–485. doi:10.1192/bjp.154.4.478

30. Ben-Zeev D, Scherer EA, Wang R, Xie H, Campbell AT. Next-generation psychiatric assessment: using smartphone sensors to monitor behavior and mental health. *Psychiatr Rehabil J*. 2015;38(3):218. doi:10.1037/prj0000130

31. Palmius N, Tsanas A, Saunders KE, et al. Detecting bipolar depression from geographic location data. *IEEE Trans Biomed Eng*. 2016;64(8):1761–1771. doi:10.1109/TBME.2016.2611862

32. Saeb S, Lattie EG, Schueller SM, Kording KP, Mohr DC. The relationship between mobile phone location sensor data and depressive symptom severity. *PeerJ*. 2016;4:e2537. doi:10.7717/peerj.2537

33. Abdullah S, Choudhury T. Sensing technologies for monitoring serious mental illnesses. *IEEE Multimed*. 2018;25(1):61–75. doi:10.1109/MMUL.2018.011921236

34. Garcia-Ceja E, Osmani V, Mayora O. Automatic stress detection in working environments from smartphones' accelerometer data: a first step. *IEEE J Biomed Health Inform*. 2015;20(4):1053–1060. doi:10.1109/JBHI.2015.2446195

35. Grünerbl A, Muaremi A, Osmani V, et al. Smartphone-based recognition of states and state changes in bipolar disorder patients. *IEEE J Biomed Health Inform*. 2014;19(1):140–148. doi:10.1109/JBHI.2014.2343154

36. Harari GM, Lane ND, Wang R, Crosier BS, Campbell AT, Gosling SD. Using smartphones to collect behavioral data in psychological science: opportunities, practical considerations, and challenges. *Perspect Psychol Sci*. 2016;11(6):838–854. doi:10.1177/1745691616650285

37. Mohr DC, Zhang M, Schueller SM. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. *Annu Rev Clin Psychol*. 2017;13:23–47. doi:10.1146/annurev-clinpsy-032816-044949

38. Wang R, Wang W, DaSilva A, et al. Tracking depression dynamics in college students using mobile phone and wearable sensing. *Proc ACM Interact Mob Wearable Ubiquitous Technol*. 2018;2(1):1–26.

39. Osmani V. Smartphones in mental health: detecting depressive and manic episodes. *IEEE Pervasive Comput*. 2015;14(3):10–13. doi:10.1109/MPRV.2015.54

40. Kim J-Y, Liu N, Tan H-X, Chu C-H. Unobtrusive monitoring to detect depression for elderly with chronic illnesses. *IEEE Sens J*. 2017;17(17):5694–5704. doi:10.1109/JSEN.2017.2729594

41. Karri SR, Khairkar P, Reddy VV. Validity of diagnostic and self-screening smartphone applications for major depressive disorders. Paper presented at: Indian Journal of Psychiatry; 2020.

42. Pratap A, Atkins DC, Renn BN, et al. The accuracy of passive phone sensors in predicting daily mood. *Depress Anxiety*. 2019;36(1):72–81. doi:10.1002/da.22822

43. Tlachac M, Rundensteiner E. Screening for depression with retrospectively harvested private versus public text. *IEEE J Biomed Health Inform*. 2020;24(11):3326–3332. doi:10.1109/JBHI.2020.2983035

44. 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. doi:10.1046/j.1525-1497.2001.016009606.x

45. Ell K, Unützer J, Aranda M, Sanchez K, Lee P-J. Routine PHQ-9 depression screening in home health care: depression prevalence, clinical and treatment characteristics, and screening implementation. *Home Health Care Serv Q*. 2006;24(4):1–19. doi:10.1300/J027v24n04_01

46. Löwe B, Schenkel I, Carney-Doebbeling C, Göbel C. Responsiveness of the PHQ-9 to psychopharmacological depression treatment. *Psychosomatics*. 2006;47(1):62–67. doi:10.1176/appi.psy.47.1.62

47. Manea L, Gilbody S, McMillan D. A diagnostic meta-analysis of the Patient Health Questionnaire-9 (PHQ-9) algorithm scoring method as a screen for depression. *Gen Hosp Psychiatry*. 2015;37(1):67–75. doi:10.1016/j.genhosppsych.2014.09.009

48. Martin A, Rief W, Klaiberg A, Braehler E. Validity of the brief patient health questionnaire mood scale (PHQ-9) in the general population. *Gen Hosp Psychiatry*. 2006;28(1):71–77. doi:10.1016/j.genhosppsych.2005.07.003

49. Nease DE, Malouin JM. Depression screening: a practical strategy. *J Fam Pract*. 2003;52(2):118–126.

50. Duffy FF, Chung H, Trivedi M, Rae DS, Regier DA, Katzelnick DJ. Systematic use of patient-rated depression severity monitoring: is it helpful and feasible in clinical psychiatry? *Psychiatr Serv.* 2008;59(10):1148–1154. doi:10.1176/ps.2008. 59.10.1148

51. 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(11):1596–1602. doi:10.1007/s11606-007-0333-y

52. Narziev N, Goh H, Toshnazarov K, Lee SA, Chung K-M, Noh Y. STDD: short-term depression detection with passive sensing. *Sensors.* 2020;20(5):1396. doi:10.3390/s20051396

53. Burchert S, Kerber A, Zimmermann J, Knaevelsrud C. Screening accuracy of a 14-day smartphone ambulatory assessment of depression symptoms and mood dynamics in a general population sample: comparison with the PHQ-9 depression screening. *PLoS One.* 2021;16(1):e0244955. doi:10.1371/journal.pone.0244955

54. Dogrucu A, Perucic A, Isaro A, et al. Moodable: on feasibility of instantaneous depression assessment using machine learning on voice samples with retrospectively harvested smartphone and social media data. *Smart Health.* 2020;17:100118. doi:10.1016/j. smhl.2020.100118

55. Schueller SM, Begale M, Penedo FJ, Mohr DC. Purple: a modular system for developing and deploying behavioral intervention technologies. *J Med Internet Res.* 2014;16(7):e181. doi:10.2196/jmir.3376

56. Spitzer RL, Kroenke K, Williams JB; Group PHQPCS. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. *JAMA.* 1999;282(18):1737–1744. doi:10. 1001/jama.282.18.1737

57. Bai R, Xiao L, Guo Y, et al. Tracking and monitoring mood stability of patients with major depressive disorder by machine learning models using passive digital data: prospective naturalistic multicenter study. *JMIR mHealth uHealth.* 2021;9(3):e24365. doi:10.2196/24365

58. Masud MT, Mamun MA, Thapa K, Lee D, Griffiths MD, Yang S-H. Unobtrusive monitoring of behavior and movement patterns to detect clinical depression severity level via smartphone. *J Biomed Inform.* 2020;103:103371. doi:10.1016/j. jbi.2019.103371

59. McIntyre RS, Lee Y, Rong C, et al. Ecological momentary assessment of depressive symptoms using the mind. me application: convergence with the Patient Health Questionnaire-9 (PHQ-9). *J Psychiatr Res.* 2021;135:311–317. doi:10.1016/j.jpsychires.2021.01.012

60. Fazel S, O'Reilly L. Machine learning for suicide research–can it improve risk factor identification? *JAMA Psychiatry.* 2020;77 (1):13–14. doi:10.1001/jamapsychiatry.2019.2896

61. Ryu S, Lee H, Lee D-K, Park K. Use of a machine learning algorithm to predict individuals with suicide ideation in the general population. *Psychiatry Investig.* 2018;15(11):1030. doi:10.30 773/pi.2018.08.27

62. Dwyer DB, Falkai P, Koutsouleris N. Machine learning approaches for clinical psychology and psychiatry. *Annu Rev Clin Psychol.* 2018;14:91–118. doi:10.1146/annurev-clinpsy-032 816-045037

63. Menton WH. Generalizability of statistical prediction from psychological assessment data: an investigation with the MMPI-2-RF. *Psychol Assess.* 2020;32(5):473. doi:10.1037/pas 0000808

64. Gradus JL, King MW, Galatzer-Levy I, Street AE. Gender differences in machine learning models of trauma and suicidal ideation in veterans of the Iraq and Afghanistan Wars. *J Trauma Stress.* 2017;30(4):362–371. doi:10.1002/jts.22210

65. Linthicum KP, Schafer KM, Ribeiro JD. Machine learning in suicide science: applications and ethics. *Behav Sci Law.* 2019;37 (3):214–222. doi:10.1002/bsl.2392

66. Mazza C, Monaro M, Orrù G, et al. Introducing machine learning to detect personality faking-good in a male sample: a new model based on Minnesota multiphasic personality inventory-2 restructured form scales and reaction times. *Front Psychiatry.* 2019;10:389. doi:10.3389/fpsyt.2019.00389

67. Oh J, Yun K, Hwang J-H, Chae J-H. Classification of suicide attempts through a machine learning algorithm based on multiple systemic psychiatric scales. *Front Psychiatry.* 2017;8:192. doi:10.3389/fpsyt.2017.00192

68. Orrù G, Mazza C, Monaro M, Ferracuti S, Sartori G, Roma P. The development of a short version of the SIMS using machine learning to detect feigning in forensic assessment. *Psychol Inj Law.* 2021;14(1):46–57. doi:10.1007/s12207-020-09389-4

69. Passos IC, Mwangi B, Cao B, et al. Identifying a clinical signature of suicidality among patients with mood disorders: a pilot study using a machine learning approach. *J Affect Disord.* 2016;193:109–116. doi:10.1016/j.jad.2015.12.066

70. Kim S, Lee H-K, Lee K. Which PHQ-9 items can effectively screen for suicide? Machine learning approaches. *Int J Environ Res Public Health.* 2021;18(7):3339. doi:10.3390/ijerph18073339

71. Piette JD, Sussman JB, Pfeiffer PN, Silveira MJ, Singh S, Lavieri MS. Maximizing the value of mobile health monitoring by avoiding redundant patient reports: prediction of depression-related symptoms and adherence problems in automated health assessment services. *J Med Internet Res.* 2013;15 (7):e2582. doi:10.2196/jmir.2582

72. Moher D, Liberati A, Tetzlaff J, Altman DG; Prisma Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med.* 2009;6(7): e1000097. doi:10.1371/journal.pmed.1000097

73. Swets JA. Measuring the accuracy of diagnostic systems. *Science.* 1988;240(4857):1285–1293. doi:10.1126/science.3287615

74. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol.* 2005;58(10):982–990. doi:10.1016/j.jclinepi. 2005.02.022

75. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med.* 2001;20(19):2865–2884. doi:10.1002/sim.942

76. Fagan TJ. Nomogram for Bayes theorem. *N Engl J Med.* 1975;293(5):257.

77. Rabe-Hesketh S, Skrondal A. *Multilevel and Longitudinal Modeling Using Stata.* STATA press; 2008.

78. Dwamena BA. *Midas: A Program for Meta-Analytical Integration of Diagnostic Accuracy Studies in Stata.* Ann Arbor: Division of Nuclear Medicine, Department of Radiology, University of Michigan Medical School; 2007.

79. Sharp S. Meta-analysis regression. *Stata Tech Bulletin.* 1998;7 (42). Available from: https://EconPapers.repec.org/RePEc:tsj: stbull:y:1998:v:7:i:42:sbe23.

80. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med.* 2011;155(8):529–536. doi:10.7326/ 0003-4819-155-8-201110180-00009

81. Breiman L. Random forests. *Mach Learn.* 2001;45(1):5–32. doi:10.1023/A:1010933404324

82. Belgiu M, Drăguţ L. Random forest in remote sensing: a review of applications and future directions. *ISPRS J Photogramm Remote Sens.* 2016;114:24–31. doi:10.1016/j.isprsjprs.2016.01.011

83. Islam MM, Iqbal H, Haque MR, Hasan MK. Prediction of breast cancer using support vector machine and K-Nearest neighbors. Paper presented at: 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC); 2017.

84. Krogh A. What are artificial neural networks? *Nat Biotechnol.* 2008;26(2):195–197. doi:10.1038/nbt1386

85. Zou J, Han Y, So -S-S. *Overview of Artificial Neural Networks*. Artificial Neural Networks; 2008:14–22.

86. Fushiki T. Estimation of prediction error by using K-fold cross-validation. *Stat Comput*. 2011;21(2):137–146. doi:10.1007/s11222-009-9153-8

87. Petitti DB. Approaches to heterogeneity in meta-analysis. *Stat Med*. 2001;20(23):3625–3633. doi:10.1002/sim.1091

88. Lau J, Ioannidis JP, Schmid CH. Summing up evidence: one answer is not always enough. *Lancet*. 1998;351(9096):123–127.

89. Sidik K, Jonkman JN. A simple confidence interval for meta-analysis. *Stat Med*. 2002;21(21):3153–3159.

90. Knapp G, Hartung J. Improved tests for a random effects meta-regression with a single covariate. *Stat Med*. 2003;22(17):2693–2710. doi:10.1002/sim.1482

91. Joyce J. *Bayes' Theorem*; The Stanford Encyclopedia of Philosophy. Edward N. Zalta(ed.) 2003. Available from: https://plato.stanford.edu/archives/fall2021/entries/bayes-theorem.

92. Simon GE, VonKorff M. Recognition, management, and outcomes of depression in primary care. *Arch Fam Med*. 1995;4(2):99–105. doi:10.1001/archfami.4.2.99

93. Shah A. The burden of psychiatric disorder in primary care. *Int Rev Psychiatry*. 1992;4(3–4):243–250. doi:10.3109/09540269209066324

94. Mitchell AJ, Yadegarfar M, Gill J, Stubbs B. Case finding and screening clinical utility of the Patient Health Questionnaire (PHQ-9 and PHQ-2) for depression in primary care: a diagnostic meta-analysis of 40 studies. *BJPsych Open*. 2016;2(2):127–138. doi:10.1192/bjpo.bp.115.001685

95. Benedetti A, Levis B, Rücker G, et al. An empirical comparison of three methods for multiple cutoff diagnostic test meta-analysis of the Patient Health Questionnaire-9 (PHQ-9) depression screening tool using published data vs individual level data. *Res Synth Methods*. 2020;11(6):833–848. doi:10.1002/jrsm.1443

96. Levis B, Benedetti A, Thombs BD. Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis. *BMJ*. 2019;365:l1476.

97. Levis B, Sun Y, He C, et al. Accuracy of the PHQ-2 alone and in combination with the PHQ-9 for screening to detect major depression: systematic review and meta-analysis. *JAMA*. 2020;323(22):2290–2300. doi:10.1001/jama.2020.6504

98. Manea L, Boehnke JR, Gilbody S, Moriarty AS, McMillan D. Are there researcher allegiance effects in diagnostic validation studies of the PHQ-9? A systematic review and meta-analysis. *BMJ Open*. 2017;7(9):e015247. doi:10.1136/bmjopen-2016-015247

99. Manea L, Gilbody S, McMillan D. Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): a meta-analysis. *Cmaj*. 2012;184(3):E191–E196. doi:10.1503/cmaj.110829

100. Fann JR, Berry DL, Wolpin S, et al. Depression screening using the Patient Health Questionnaire-9 administered on a touch screen computer. *Psycho Oncol*. 2009;18(1):14–22. doi:10.1002/pon.1368

101. Spitzer RL, Williams JB, Kroenke K, Hornyak R, McMurray J; Group PHQO-GS. Validity and utility of the PRIME-MD patient health questionnaire in assessment of 3000 obstetric-gynecologic patients: the PRIME-MD Patient Health Questionnaire Obstetrics-Gynecology Study. *Am J Obstet Gynecol*. 2000;183(3):759–769. doi:10.1067/mob.2000.106580

102. Alfonsson S, Maathz P, Hursti T. Interformat reliability of digital psychiatric self-report questionnaires: a systematic review. *J Med Internet Res*. 2014;16(12):e268. doi:10.2196/jmir.3395

103. Erbe D, Eichert H-C, Rietz C, Ebert D. Interformat reliability of the patient health questionnaire: validation of the computerized version of the PHQ-9. *Internet Interv*. 2016;5:1–4. doi:10.1016/j.invent.2016.06.006

104. Zhen L, Wang G, Xu G, et al. Evaluation of the paper and smartphone versions of the Quick Inventory of Depressive Symptomatology-Self-Report (QIDS-SR16) and the Patient Health Questionnaire-9 (PHQ-9) in depressed patients in China. *Neuropsychiatr Dis Treat*. 2020;16:993. doi:10.2147/NDT.S241766

105. Streiner DL, Norman GR, Cairney J. *Health Measurement Scales: A Practical Guide to Their Development and Use*. USA: Oxford University Press; 2015.

106. Williams JW Jr, Noël PH, Cordes JA, Ramirez G, Pignone M. Is this patient clinically depressed? *JAMA*. 2002;287(9):1160–1170. doi:10.1001/jama.287.9.1160

107. Jin H, Wu S. Text messaging as a screening tool for depression and related conditions in underserved, predominantly minority safety net primary care patients: validity study. *J Med Internet Res*. 2020;22(3):e17282. doi:10.2196/17282

108. Ware S, Yue C, Morillo R, et al. Predicting depressive symptoms using smartphone data. *Smart Health*. 2020;15:100093. doi:10.1016/j.smhl.2019.100093

109. Lawson A, Dalfen A, Murphy KE, Milligan N, Lancee W. Use of text messaging for postpartum depression screening and information provision. *Psychiatr Serv*. 2019;70(5):389–395. doi:10.1176/appi.ps.201800269

110. McGinnis RS, McGinnis EW, Hruschak J, et al. Rapid anxiety and depression diagnosis in young children enabled by wearable sensors and machine learning. Paper presented at: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2018.

111. Pfeiffer PN, Bohnert KM, Zivin K, et al. Mobile health monitoring to characterize depression symptom trajectories in primary care. *J Affect Disord*. 2015;174:281–286. doi:10.1016/j.jad.2014.11.040

112. Mastoras R-E, Iakovakis D, Hadjidimitriou S, et al. Touchscreen typing pattern analysis for remote detection of the depressive tendency. *Sci Rep*. 2019;9(1):1–12. doi:10.1038/s41598-019-50002-9

113. Razavi R, Gharipour A, Gharipour M. Depression screening using mobile phone usage metadata: a machine learning approach. *J Am Med Inform Assoc*. 2020;27(4):522–530. doi:10.1093/jamia/ocz221

114. Bakker D, Rickard N. Engagement in mobile phone app for self-monitoring of emotional wellbeing predicts changes in mental health: moodPrism. *J Affect Disord*. 2018;227:432–442. doi:10.1016/j.jad.2017.11.016

115. BinDhim NF, Alanazi EM, Aljadhey H, et al. Does a mobile phone depression-screening app motivate mobile phone users with high depressive symptoms to seek a health care professional's help? *J Med Internet Res*. 2016;18(6):e5726. doi:10.2196/jmir.5726

116. Kim H, Lee S, Lee S, Hong S, Kang H, Kim N. Depression prediction by using ecological momentary assessment, actiwatch data, and machine learning: observational study on older adults living alone. *JMIR mHealth uHealth*. 2019;7(10):e14149. doi:10.2196/14149

117. Valstar M, Gratch J, Schuller B, et al. Avec 2016: depression, mood, and emotion recognition workshop and challenge. Paper presented at: Proceedings of the 6th international workshop on audio/visual emotion challenge; 2016.

118. Gradus JL, Rosellini AJ, Horváth-Puhó E, et al. Prediction of sex-specific suicide risk using machine learning and single-payer health care registry data from Denmark. *JAMA Psychiatry*. 2020;77(1):25–34. doi:10.1001/jamapsychiatry.2019.2905

119. Chekroud AM, Bondar J, Delgadillo J, et al. The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry*. 2021;20(2):154–170. doi:10.1002/wps.20882

120. Pearson SD, Katzelnick DJ, Simon GE, Manning WG, Helstad CP, Henk HJ. Depression among high utilizers of medical care. *J Gen Intern Med*. 1999;14(8):461–468. doi:10.1046/j.1525-1497.1999.06278.x

121. Almeida OP. Prevention of depression in older age. *Maturitas*. 2014;79(2):136–141. doi:10.1016/j.maturitas.2014.03.005

122. Hall CA, Reynolds-Iii CF. Late-life depression in the primary care setting: challenges, collaborative care, and prevention. *Maturitas*. 2014;79(2):147–152. doi:10.1016/j.maturitas.2014.05.026

123. Hegadoren K, Norris C, Lasiuk G, Silva D, Chivers-Wilson K. The many faces of depression in primary care. *Texto Contexto Enferm*. 2009;18:155–164. doi:10.1590/S0104-07072009000100019

124. Park LT, Zarate CA Jr. Depression in the primary care setting. *N Engl J Med*. 2019;380(6):559–568. doi:10.1056/NEJMcp1712493

125. Valenstein M, Vijan S, Zeber JE, Boehm K, Buttar A. The cost–utility of screening for depression in primary care. *Ann Intern Med*. 2001;134(5):345–360. doi:10.7326/0003-4819-134-5-200103060-00007