Journal of Affective Disorders 376 (2025) 216–224

Contents lists available at ScienceDirect

Journal of Affective Disorders

journal homepage: www.elsevier.com/locate/jad

Research paper

A machine learning model using clinical notes to estimate PHQ-9 symptom severity scores in depressed patients

Pedro Alves [a], Carl D. Marci [a,b,*], Chandra J. Cohen-Stavi [a], Katelynn Murray Whelan [c], Costas Boussios [a]

[a] OM1, Inc., Boston, MA, United States of America
[b] Massachusetts General Hospital, Harvard Medical School, Boston, MA, United States of America
[c] University College Dublin School of Medicine, Dublin, Ireland

ARTICLE INFO

Keywords:
Machine learning
Major depressive disorder
PHQ-9
Real-world data
Structure and unstructured clinical notes

ABSTRACT

Background: Lack of widespread use of the Patient Health Questionnaire 9-item (PHQ-9) in clinical practice inhibits measurement of treatment follow-up for patients with major depressive disorder (MDD). This study developed, validated and applied a machine learning model to estimate PHQ-9 scores for MDD patients using relevant notes from electronic medical records (EMR).

Methods: Information from structured and unstructured sections of prescriber notes from a multi-source real-world mental health database were used to estimate PHQ-9 scores (ePHQ-9). Model performance and agreement were evaluated using binary and categorical PHQ-9 outcomes. The final model strategy was applied to MDD patient encounters without scores to assess the extent of added available PHQ-9 measures.

Results: A final model was developed from 48,594 patients and 143,224 clinical encounters with a recorded PHQ-9 score, and then applied to 196,819 MDD patients. Overall model performance was high with an AUC 0.81, PPV 0.71 and NPV 0.76. The addition of ePHQ-9 scores increased the average number of available scores per patient per year by 2.8×.

Limitations: The model was developed using prescribing mental health providers' clinical notes, which limits generalizability to other contexts (e.g., primary care). The PHQ-9 is designed to be patient-reported, whereas this model strategy estimates PHQ-9 scores using clinicians' notes, which results in some expected discrepancies.

Conclusions: This validated ePHQ-9 model contributes to addressing measurement gaps in depression treatment and research by adding substantially to the number of measures available in real-world data for clinical follow-up.

1. Background

Major depressive disorder (MDD) is a common psychiatric condition whose health and economic burden has been rising substantially over recent years. In 2020, 18.4 % of the U.S. adult population reported a past history of depression (Lee et al., 2020). Depressive and anxiety disorders were a leading cause for concern globally before the COVID-19 pandemic, with clear evidence of increasing prevalence of those suffering from depression between 2015 and 2019 (Goodwin et al., 2022). These trends have worsened with the pandemic contributing to a growing public health crisis with staggering economic costs associated with MDD estimated at over $326 billion in recent years (Santomauro et al., 2021; Goodwin et al., 2022; Greenberg et al., 2021).

With this rising burden of depression, there has been an emphasis on implementing measurement-based care to monitor disease progression, inform treatment management and improve outcomes, but there are barriers to widespread adoption (Hong et al., 2021). Multiple attempts by researchers to identify physiologic, genetic or digital biomarkers that indicate treatment response or depression subtypes have struggled to establish broad clinical acceptance (Fraguas et al., 2007; Zeier et al., 2018; Place et al., 2020). In the absence of objective biomarkers for tracking disease status, the primary option available for depression screening, diagnosis and monitoring is through the elicitation of patient symptoms by clinical interview or patient reported measures. Multiple clinician and patient reported scales exist for assessing depression symptoms and severity, but few are used consistently in clinical practice

* Corresponding author at: 31 St James Ave #1010, Boston, MA 02116, United States of America.
E-mail address: cmarci@om1.com (C.D. Marci).

https://doi.org/10.1016/j.jad.2025.01.152
Received 17 May 2024; Received in revised form 26 January 2025; Accepted 31 January 2025
Available online 3 February 2025
0165-0327/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

(Gliklich et al., 2020; Cheung et al., 2023).

The Patient Health Questionnaire 9-item (PHQ-9), consisting of the nine depression criteria from the DSM-IV, is one of the more commonly used instruments for diagnosing and monitoring symptom severity in MDD in the real-world (Kroenke et al., 2001; Kroenke, 2021; Löwe et al., 2004). The brevity, discriminatory ability and sensitivity to detecting change in depression severity make the PHQ-9 an attractive tool to help screen, diagnose and monitor depression in different care settings (Inoue et al., 2012). Despite wide acceptance of the PHQ-9, there is limited and sporadic use in psychiatric and primary care practice, resulting in inconsistent documentation and only modest availability in real-world data sources (Ford et al., 2020; Gliklich et al., 2020).

Machine learning has been used in attempts to address some of the gaps in depression screening and monitoring. These attempts include efforts to predict antidepressant treatment outcomes based on data from electronic medical records (EMR); to identify depression severity through EMR clinical note extraction; to detect behavioral patterns of depression using smartphone data; and to predict healthcare professionals at-risk for developing symptoms of depression (Xu et al., 2023; Zimmerman and McGlinchey, 2008; Zhou et al., 2022; Adekkanattu et al., 2018; Choudhary et al., 2022). Most of these studies demonstrate the feasibility of using machine learning models for identifying or predicting depression, assessment of depression symptom severity or evaluation of depression treatment effects in relatively small samples or in a single care setting. To date, there are no models that can be applied on a larger scale to fill in the gaps in depression research and to augment clinical care.

The objectives of the current effort were to validate a machine learning model to estimate PHQ-9 scores (ePHQ-9) based on prescriber EMR notes and to apply the model to a large, multisource real-world dataset to generate ePHQ-9 scores for use in future research studies.

2. Methods

2.1. Participants

A training and validation cohort of patient data with a training and validation set of encounters for model development were drawn from a large multisource real-world mental health specialty network (OM1, Inc., Boston MA). The specialty network includes data from 2013 to present of deterministically linked, de-identified, patient-level health care information from structured and unstructured EMR notes, unadjudicated insurance and pharmacy claims and other specialty and primary care sources on over 3.5 million patients receiving treatment from over 9000 mental health professionals across 2500 clinics in all 50 United States. The model was then applied to a subset of over 450,000 patients who qualify for a condition dataset specific to MDD (OM1 MDD PremiOM™ Dataset, Boston MA) to evaluate the extent to which the estimation model adds to available outcome data for tracking and monitoring depression (the estimation cohort).

Individuals were included in the training and validation cohorts based on whether they had an encounter with a mental health prescriber with a clinical note meeting sufficient data requirements of length and complexity (defined below). The model was not limited by study population to those with an MDD diagnosis because the PHQ-9 is used as a screening and diagnostic tool, as well as an assessment tool. Therefore, it was important to capture patients with and without a diagnosis of MDD. Individuals with both reported PHQ-9 scores and EMR notes from mental health prescribers were identified and randomly assigned to the training cohort or the validation cohort for model development. Mental health prescribers included psychiatrists, nurse practitioners and physician assistants working within the mental health specialty network.

When applying the model to the estimation cohort to evaluate the extent to which the new ePHQ-9 scores enhance outcome data for depression care follow-up, the focus was on those patients who already had a diagnosis of MDD. This estimation cohort included patients within the OM1 MDD PremiOM™ Dataset, which requires that patients meet at least one of the following criteria: a minimum of two MDD diagnostic codes at least 30 days apart from encounters with a mental health specialist; at least one inpatient visit with an MDD diagnostic code; or at least two outpatient records with an MDD diagnostic code at least 30 days apart and within a year, regardless of clinician specialty. Among this population, the model application was limited to EMR notes with sufficient data requirements (defined below) from mental health prescribers.

2.2. Model development: dependent variable

The models were trained to estimate PHQ-9 scores, which are based on nine questions with each individual response ranging from 0 to 3 resulting in a total score from 0 to 27. The trained model generates an estimated PHQ-9 score (ePHQ-9) that also ranges from 0 to 27 using information from a clinical encounter on a specific date. Because scores are estimated for a clinical encounter on a specific date, patients often have multiple ePHQ-9 scores and additional timepoints for which they have reported PHQ-9 scores.

2.3. Model development: types of models

In addition to providing space for free-text narrative note taking, modern EMR systems also commonly include elements where a clinician can click buttons or use pre-populated dropdown menus or other forms of content with structured, predetermined outputs in order to increase note keeping efficiency. Such content appears in the OM1 MDD PremiOM™ Dataset as structured content. For example, a "Review of Systems" may appear as a set of checked items for each organ system and returns a 0 for the absence and a 1 for the presence of the symptom (e.g., the presence of "back pain" is coded in a dedicated data field as "1"). The ePHQ-9 model treats the free-text, unstructured content separately from the structured content.

Three models were developed: 1) an unstructured clinical notes based model using unstructured narrative text; 2) a structured numeric data model using numeric inputs from templates from the EMR; and 3) an ensemble model that combines the first two models. This approach allows the incorporation of diverse EMR data types and increases model performance. A training and validation cohort was defined for each of the three models developed and there was no patient overlap across cohorts. The combined training cohort for all three models included 67.5 % of available patients and the combined validation cohort 32.5 % of available patients.

Relevant data types were identified based on fields from clinical encounter EMR records, including relevant unstructured clinical text and structured numeric data. To prevent overfitting for patients with many encounters in their EMR, the maximum number of encounters was limited to ten per patient (when the limit was enforced, the encounters were selected by random sampling). To be eligible for inclusion in the modeling, the unstructured clinical notes and structured numeric data were required to have sufficient requirements defined as:

1. Unstructured clinical notes were required to have data in at least three of four note sections including "History of Present Illness," "History of Present Illness/Interval History," "Chief Complaint," and "Assessment/Impression/Plan."
2. Structured numeric data from the notes was required to have data in the "Review of Systems" section of the record and documentation of at least one of the following structured sections related to suicide assessment: "Stressors," "Suicide Prevention Plan," "Suicide Protection Factors," or "Suicide Risk Factors."

2.4. Model development: explanatory variables

The explanatory variables (predictive features of the model) were derived from multiple sources of information. The unstructured clinical notes based model relied on language patterns found in unstructured free text data contained in the EMR notes. In the unstructured model, the clinical notes were transformed via medical language processing including steps to remove auto-templating and propagated negations. Explanatory variables were derived from terms and expressions including those indicating clinical progress (e.g., improvement, stability, or worsening); the presence or absence of comorbidities (e.g., substance use, anxiety disorders); and the presence or negation of relevant signs and symptoms (e.g., mood changes, weight changes, eating habits, level of concentration, sleep problems, energy level). Predictive features also included some items similar to some of the PHQ-9 questions (e.g., levels of fatigue or the presence of suicidal ideation) as well as other clinically relevant concepts (e.g., medication usage and condition-specific scores). The structured numeric data model utilizes structured numeric translations of answers provided on questionnaires that utilize templates covering areas such as review of systems, medication side effects, stressors (e.g., housing, economic, social support, occupational status, educational level) and other clinically relevant indicators (e.g., suicide risk factors and presence of a suicide prevention plan). The predictive features of both the structured and unstructured models were reviewed for clinical relevance by a clinician expert. The inputs for the final ensemble model include information from both the unstructured clinical notes and the structured numeric data models.

2.5. Model strategy

All three models (unstructured clinical notes, structured numeric data, and ensemble) employed the XGBoost algorithm. Logistic regression and random forest algorithms were also evaluated. For each eligible encounter, the ePHQ-9 is based on the following logic to combine the three models: if the sufficient data requirements are met for both the unstructured clinical notes model and the structured numeric data model, then the ePHQ-9 is estimated using the ensemble model output. If the data requirements are met by only the unstructured clinical notes model, then the output of that model is used for ePHQ-9 prediction. The structured numeric data model was deemed insufficient alone and was not used by itself to produce ePHQ-9 scores. The final model logic is outlined in Fig. 1.

To assess the overall model performance, the area under the receiver-operating characteristic curve (AUC), positive predictive value (PPV) and negative predictive value (NPV) were calculated for the final model using a binary version of the outcome, with a threshold PHQ-9 score of 10 representing moderate to severe levels of depression were used with "high" defined as PHQ-9 scores greater than or equal to 10 and "low" defined as all other scores. Evaluation of continuous ePHQ-9 scores compared to reported PHQ-9 scores was conducted by calculating Spearman R and Pearson R values. The distribution of ePHQ-9 scores was also compared to the distribution of reported PHQ-9 scores. Two confusion matrices were generated. The first compares the ePHQ-9 and reported PHQ-9 scores in the validation set for the binary PHQ-9 outcome (scores <10 versus scores greater than or equal to 10). The second was based on more detailed clinical categorization of depression severity based on PHQ-9 scores (Minimal: 0–4; Mild: 5–9; Moderate: 10–14; Moderately Severe: 15–19; Severe: 20–27).

2.6. Application of the model

The final model strategy was applied to a set of MDD patient encounter notes in the estimation cohort to generate ePHQ-9 scores that met the clinical notes data sufficiency requirements but did not have reported PHQ-9 scores. For ePHQ-9 scores in the overall validation set, isotonic regression was used to calibrate the ePHQ-9 to the reported

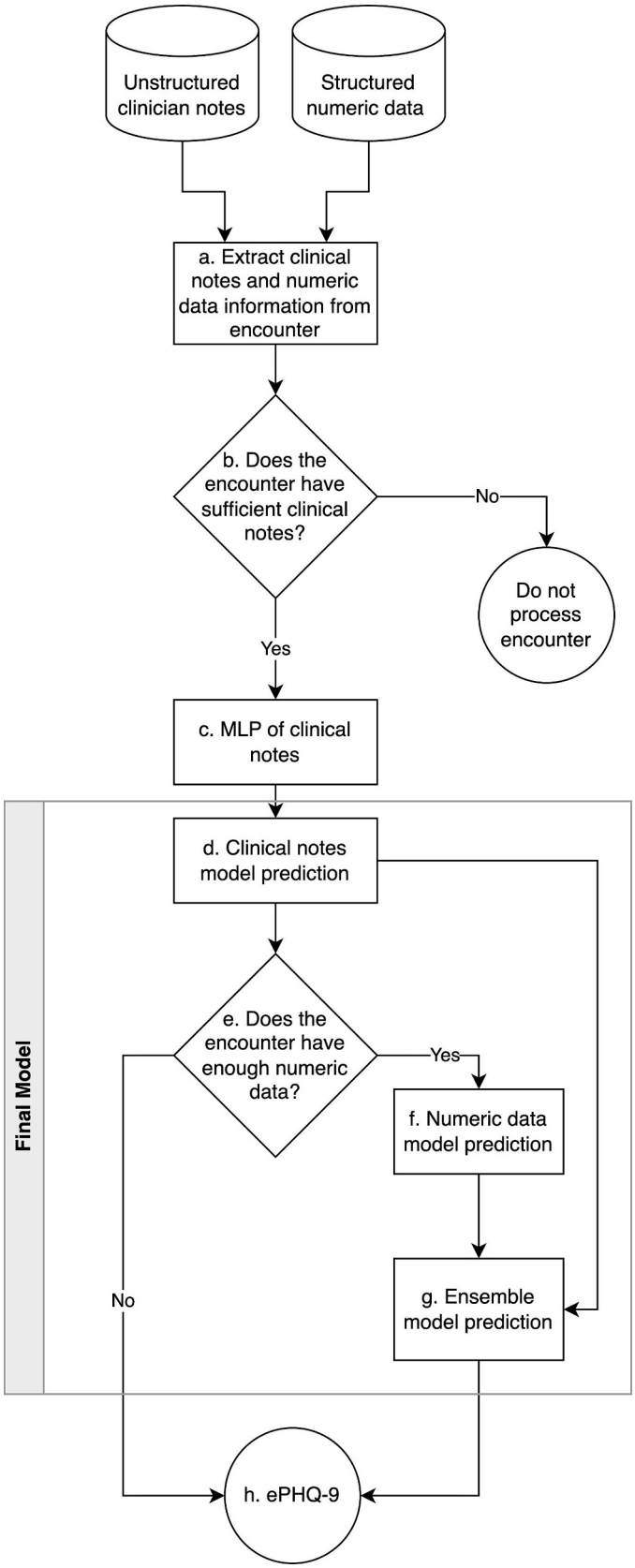

Fig. 1. Final model logic based on data elements available in the relevant clinical notes.

PHQ-9 scores, which yielded lower volatility than other calibration methods (De Leeuw et al., 2010). The distribution of ePHQ-9 scores in the estimated set was compared to the reported PHQ-9 scores in the validation set. Finally, to assess the extent to which the modeled ePHQ-9 scores fill gaps in MDD patient care journeys, the average and median number of reported and ePHQ-9 scores available per patient per year were calculated among patients in the MDD estimation cohort. The eligible period per patient was defined as the time between the first year a patient has an encounter in the database through 2023. The average and median number of PHQ-9 scores were calculated per year per patient over the eligible period. These were then aggregated at the study population level by computing the average of the mean and median scores per year over all patients.

3. Results

The model development cohort consisted of 48,594 patients with a total of 143,224 clinical encounters that had a recorded PHQ-9 score. This cohort was divided into a training cohort and a validation cohort. The training cohort consisted of 32,802 patients with a total of 96,891 encounters with a recorded PHQ-9 score. This set of encounters was the training set for the models. The validation cohort consisted of 15,792 patients with a total of 46,333 encounters with a recorded PHQ-9 score, which constituted the validation set. The training cohort was used for the training of all three models (unstructured clinical notes, structured numeric data and ensemble). Of note, 99.8 % of patients had at least one diagnosis of MDD in the training and validation cohorts.

Table 1 compares characteristics of the training, validation and estimation cohort. Since many patients have both recorded and ePHQ-9 scores, the estimation cohort has patients in common with the model development cohorts. In contrast, the training and validation cohort have no patients in common. The demographic characteristics of the training and validation cohorts were similar with a mean age of 38.4 and 38.5 years, 67.3 % and 67.0 % female, between 44.1 % and 44.7 % white race, respectively. Among the 196,819 individuals in the MDD estimation cohort, the mean age was slightly higher at 41.0 years but with similar demographic characteristics. The average number of reported PHQ-9 scores were 1.56 (SD = 1.34) in the training cohort, 1.57 (SD = 1.42) in the validation cohort and 1.42 (SD = 1.31) in the estimation cohort (Table 1).

The final model had an AUC of 0.81, a PPV of 0.71 and a NPV of 0.76 when evaluating performance in the validation set using the binary outcome (Fig. 2). Performance evaluated using the continuous ePHQ-9 scores yielded a Spearman R value of 0.62 and a Pearson R value of 0.61. Two confusion matrices were generated to further assess the agreement between the ePHQ-9 scores and recorded PHQ-9 scores (see Fig. 3a for binarized scores and Fig. 3b for categories of PHQ-9 and ePHQ-9). Fig. 4 shows the comparison of the distribution of reported PHQ-9 scores to the distribution of ePHQ-9 scores in the validation cohort.

In the estimation cohort, ePHQ-9 scores for 2,113,646 distinct encounters from 196,819 patients without a reported PHQ-9 were derived based on the application of the model. A comparison between the distribution of reported PHQ-9 scores in the validation cohort and ePHQ-9 scores in the estimation cohort yielded more estimated scores on the lower end of the scale (Fig. 5).

Table 2 shows that the addition of the ePHQ-9 scores increased the average number of available scores per patient per year by a factor of 2.8×. Of 528,041 patients in the MDD dataset, the mean number of reported PHQ-9 scores per year per patient was 0.42 (SD = 1.32) while the mean number of combined reported or ePHQ-9 scores per patient per year increased to 1.18 (SD = 2.39). The aggregated average of the median number of observed PHQ-9 scores per patient per year was 0.29 (SD = 1.26) and with combined observed PHQ-9 or ePHQ-9 scores, the average of the median was 0.89 (SD = 2.38) with an increase of a factor of >3×. Almost two-thirds of the MDD patients (63.4 %) had no reported

Table 1
Comparison of demographic, data and clinical characteristics of the training, validation and estimation cohorts.

Cohort characteristics		Training cohort	Validation cohort	Estimation cohort
N		32,802	15,792	196,819
Demographic characteristics				
Age at Date of First Reported PHQ-9 Score, Mean (SD)		38.4 (17.6)	38.5 (17.5)	41.0 (16.9)
Sex	Female	22,083 (67.3 %)	10,744 (68.0 %)	134,836 (68.5 %)
	Male	10,704 (32.6 %)	5042 (31.9 %)	61,983 (31.5 %)
	Unknown	15 (0.05 %)	6 (0.04 %)	0 (0.00 %)
Race	Asian	244 (0.74 %)	116 (0.73 %)	1610 (0.82 %)
	White	14,453 (44.06 %)	7053 (44.66 %)	78,550 (39.91 %)
	Black	1108 (3.38 %)	506 (3.20 %)	6162 (3.13 %)
	Other	660 (2.01 %)	307 (1.94 %)	3780 (1.92 %)
	Unknown	16,337 (49.80 %)	7810 (49.46 %)	106,717 (54.22 %)
Census region	Northeast	5637 (17.18 %)	2630 (16.65 %)	39,580 (20.11 %)
	Midwest	5137 (15.66 %)	2460 (15.58 %)	33,214 (16.88 %)
	South	10,034 (30.59 %)	4937 (31.26 %)	68,408 (34.76 %)
	West	11,952 (36.44 %)	5744 (36.37 %)	55,434 (28.16 %)
	Unknown	42 (0.13 %)	21 (0.13 %)	183 (0.09 %)
Data-related characteristics				
Years of follow-up in database, Mean (SD)		0.78 (1.44)	0.79 (1.47)	0.62 (1.32)
Number of encounters,[a] Mean (SD)		13.5 (28.37)	13.34 (28.96)	8.87 (21.97)
Number of reported PHQ-9 s, Mean (SD)		1.56 (1.34)	1.57 (1.42)	1.42 (1.31)
Clinical characteristics[b] - comorbidities and depression treatments				
Charlson score category	0–1	29,082 (88.7 %)	13,992 (88.6 %)	173,066 (87.9 %)
	2–3	2599 (7.9 %)	1239 (7.6 %)	16,489 (8.4 %)
	4–5	624 (1.9 %)	347 (2.2 %)	4037 (2.1 %)
	6+	497 (1.5 %)	214 (1.7 %)	3227 (1.6 %)
Anxiety		17,730 (54.1 %)	8496 (53.8 %)	88,726 (45.1 %)
ADHD		5710 (17.4 %)	2769 (17.5 %)	27,807 (14.1 %)
Substance use disorder		3296 (10.1 %)	1628 (10.3 %)	16,031 (8.2 %)
Antidepressant medications		23,181 (70.7 %)	11,083 (70.2 %)	115,038 (58.5 %)
Antidepressant + augmentation medications		5873 (17.9 %)	2826 (17.9 %)	27,629 (14.0 %)
Any psychotherapy		25,189 (76.8 %)	12,087 (76.5 %)	121,231 (61.6 %)

[a] The number of encounters was limited to the mental health specialty database at OM1, of whom a majority are MH specialists.

[b] All clinical characteristics were identified during the 6 months prior to the date of the first reported PHQ-9 score in the training and validation cohorts and during the 6 months prior to the date of the first ePHQ-9 score in the estimation cohort.

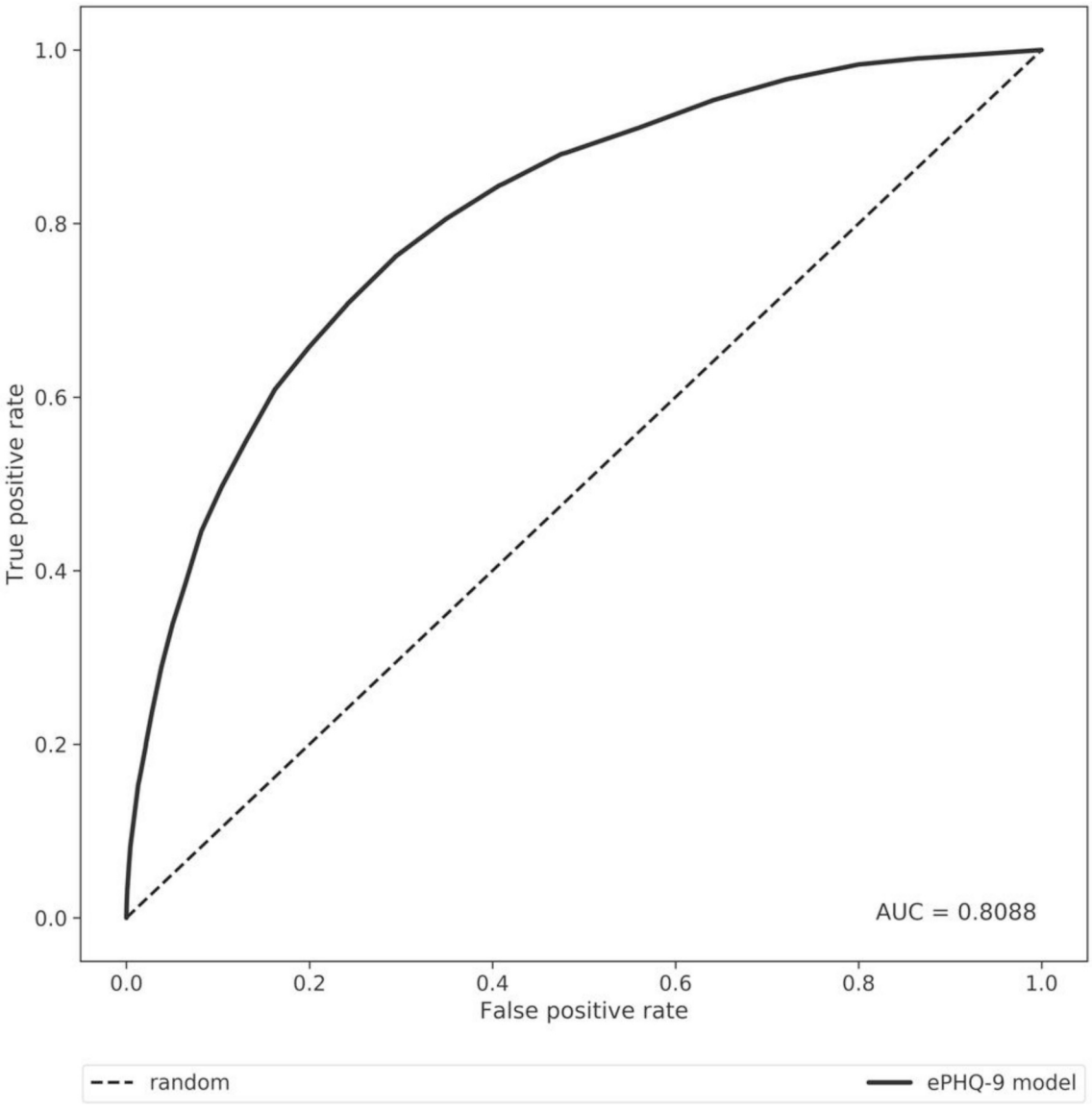

Fig. 2. Area under the receiver-operating characteristic curve for the final model.

PHQ-9 scores in the database prior to running the estimation model. Following the application of the estimation model, that proportion dropped to 43.3 %. Proportions of patients that had two, three, four, five or more total scores in the database increased substantially with the ePHQ-9 model scores.

4. Discussion

We present a machine learning model to estimate PHQ-9 scores based on structured and unstructured elements from EMR notes generated in routine clinical care of patients with documented major depressive disorder. The agreement between the reported and ePHQ-9 scores was high for classifying between mild and minimal depression (scores of 0–9) and moderate to severe depression (scores of 10–27). Agreement was lower when classifying more granular PHQ-9 score categories (i.e., minimal, mild, moderate, moderate-severe and severe) as there is a larger number of classes that are compared. The output was applied to a large real-world dataset to fill in the gaps where PHQ-9 scores were not documented. As noted in Fig. 5, more scores were estimated at the lower end of the scale which may reflect the fact that clinicians are more likely to ask patients with higher disease burden to fill out the PHQ-9. Importantly, the ePHQ-9 added substantially to the average number of available PHQ-9 scores overall and resulted in much higher proportions of patients that had two, three, four, five or more total scores in the database.

Prior studies have demonstrated the feasibility of employing machine learning based on various data sources to extract information related to depression symptoms and disease severity from clinical notes for predicting treatment outcomes or for modeling behavioral patterns related to depression (Xu et al., 2023; Zhou et al., 2015; Adekkanattu et al., 2018; Choudhary et al., 2022). The performance of the final ePHQ-9 model in the current study with an AUC = 0.81 is considered in the literature as ranging from “good” to “excellent” (White et al., 2023). In contrast, the Pearson coefficient value of 0.61 of the individual scores is considered “moderate” (Schober et al., 2021). The results from the final model are also comparable to the higher performing machine learning models previously published and developed for use in MDD that had AUCs ranging from 0.61 to 0.83 (Xu et al., 2023; Choudhary et al., 2022; Coley et al., 2021; Hatton et al., 2019). However, the previous models that were developed for predicting outcomes in depression were based predominantly on claims data and generally had lower performance than models used to estimate depression symptoms in isolation or

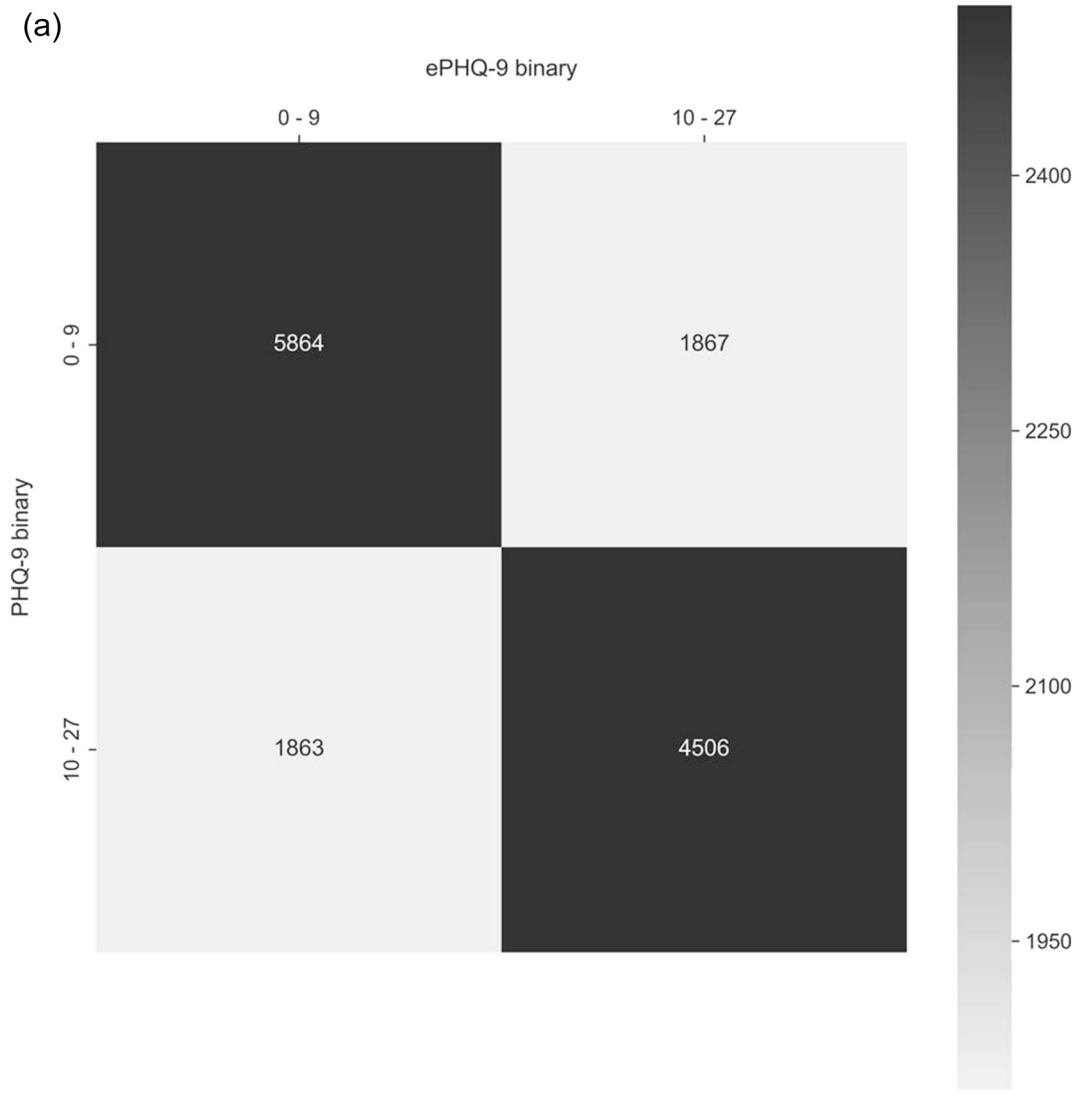

Fig. 3a. Confusion matrix of the agreement between estimated and reported PHQ-9 scores in the validation cohort based on binary low/high classes of PHQ-9 score classes.

overall disease severity based on clinical note data. In the current study, the ePHQ-9 model relies on the latter, unstructured free-text and standardized structured elements from relevant sections of the EMR notes.

The current approach has several strengths including the use of large and diverse training and validation cohorts that are geographically and demographically representative of patients seen in mental health clinics in the US. Another strength of this effort is the incorporation of both structured and unstructured data from sections of relevant EMR notes from mental health prescribers into the model development and validation. When compared with psychotherapists' notes, a preliminary review of the dataset suggested mental health prescribers more consistently documented symptoms of depression and therefore their notes were more likely to contain information needed to estimate the PHQ-9. This is important, as noted in the description of explanatory variables in the methods section, the predictive features used in the final model are clinically relevant and mirror several of the core symptoms of MDD. This should not be surprising given most prescribers use the DSM criteria for tracking disease severity and response to treatment even in the absence of clinical scales. Despite the fact that clinical notes vary significantly in the consistency of documentation, the clinical relevance of the features contributes to the explainability of the model. While some researchers have cautioned against using EMR-based prediction models for depression outcomes given the risks and drawbacks in applying them in clinical practice, the performance of the ensemble model suggests that computer based estimations may help answer calls for more widespread use of PHQ-9 scores in clinical practice (Ford et al., 2020; Gliklich et al., 2020).

There are also several limitations to consider. The estimation model for the PHQ-9 was trained on prescribing clinicians' EMR notes only, which may differ from clinical notes and observations from other providers. This may limit the generalizability of the model in other clinical contexts, such as primary care settings where large numbers of patients with depression are treated (Barkil-Oteo, 2013). Another limitation is that the PHQ-9 was designed to be patient administered. This study uses clinicians' notes to estimate PHQ-9 scores which may explain the decrease in performance of the model at more granular levels of comparison. Given the PHQ-9's versatility and use as both a screening and monitoring tool, there is likely wide variation in how it is administered and different kinds of biases can be present due to the inherent subjectivity of the instrument. Very high and very low scores more consistently represent delineations in depression symptoms and severity, while other scores in the middle range represent a spectrum of patient presentations (Kroenke, 2021). When examining the agreement between the current study's ePHQ-9 model and patient reported PHQ-9 scores, the

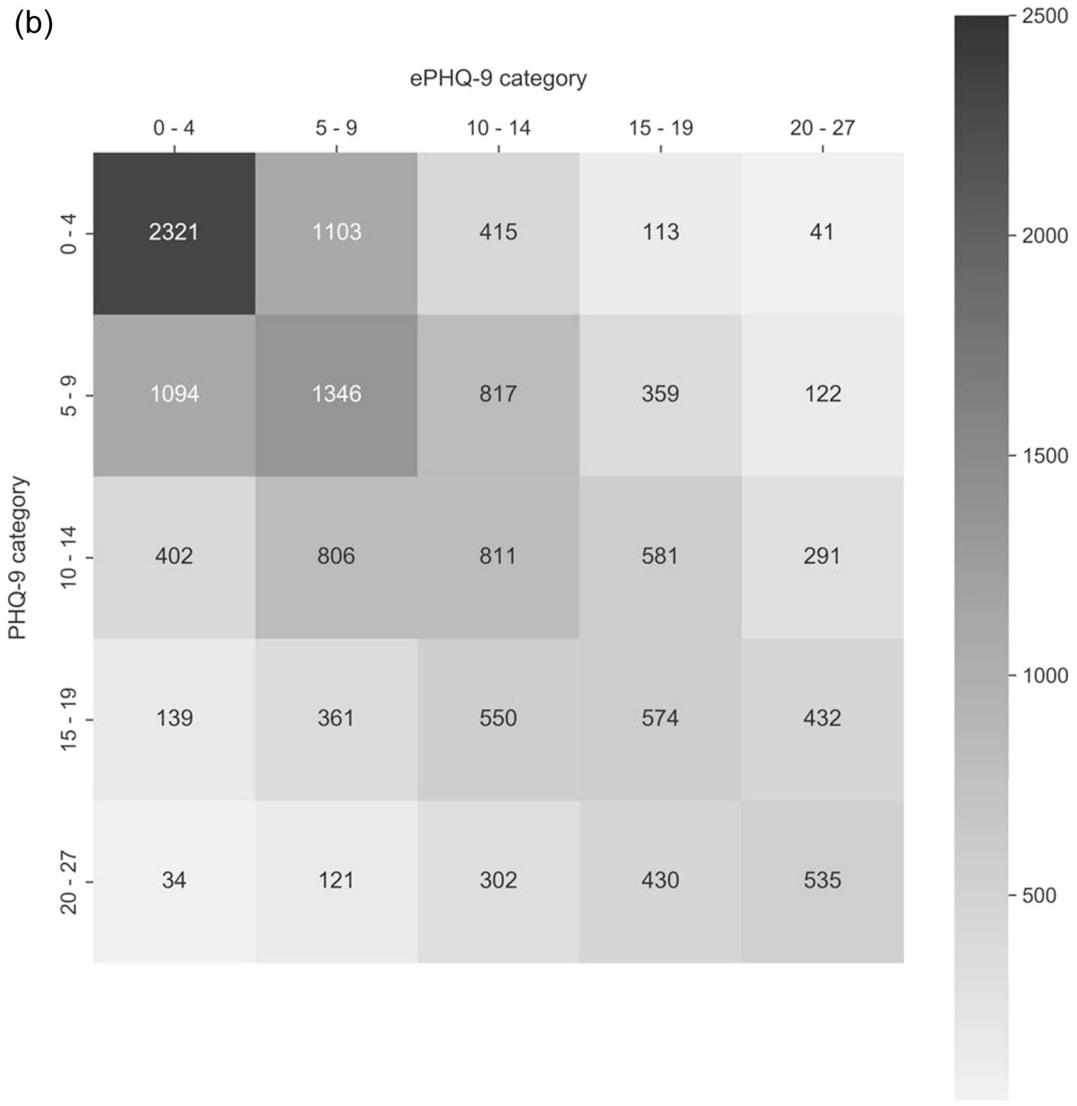

Fig. 3b. Confusion matrix of the agreement between estimated and reported PHQ-9 scores in the validation cohort based on categories of PHQ-9 scores.

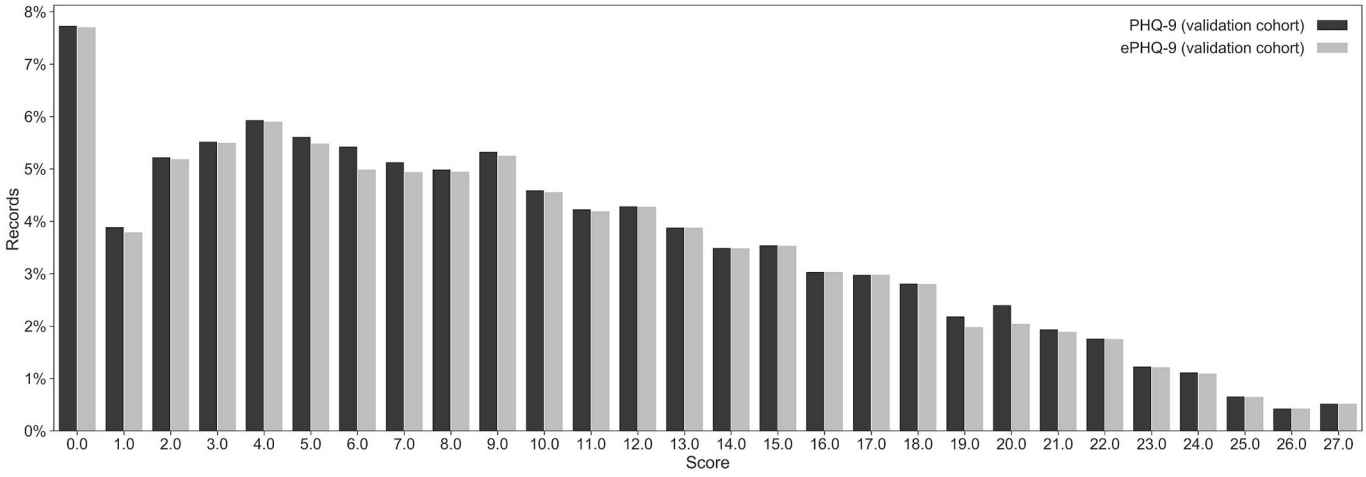

Fig. 4. Distribution of PHQ-9 scores vs ePHQ-9 scores in the validation set.

relationship varied depending on the granularity of the PHQ-9 outcome definition. This is consistent with a prior study that showed decreasing accuracy and precision as the model targeted increasing granularity in outcome categories of depression severity (Choudhary et al., 2022). Prior research has also demonstrated that there are discrepancies between observer-rated and patient-rated depression symptoms due to

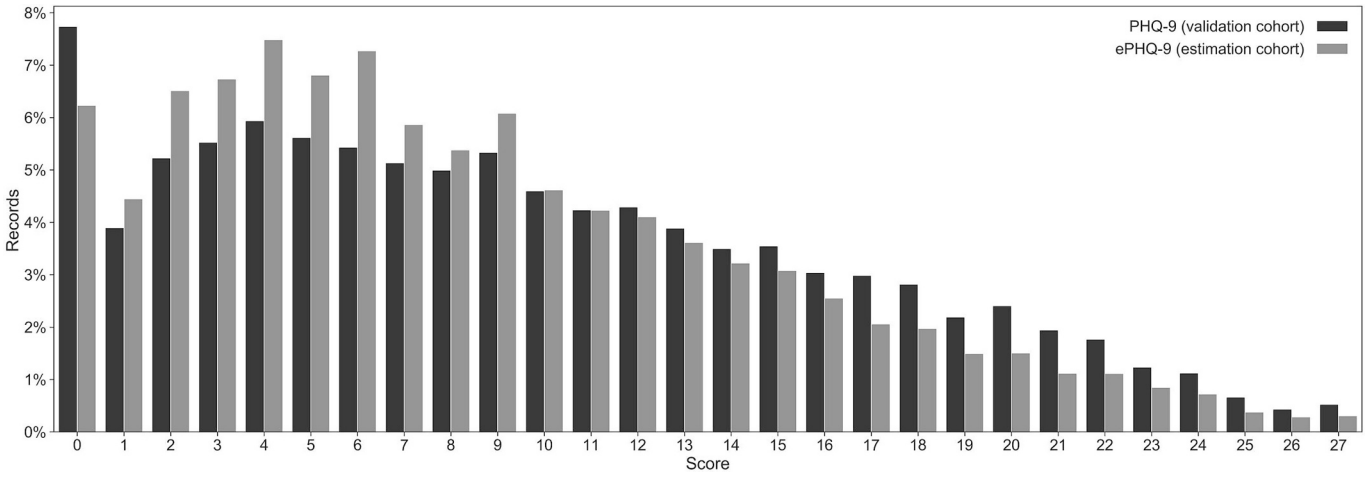

Fig. 5. Distribution of PHQ-9 scores in the validation set vs ePHQ-9 scores among patients with no PHQ-9 data.

Table 2
Average and median number of scores per patient per year and the number of total scores among patients with reported PHQ-9 scores compared to patients with either reported or estimated scores.

	PHQ-9 scores prior to estimation model	PHQ-9 scores after estimation model
Total Patients in OM1 MDD PremiOM™ Dataset, N	528,041 (100 %)	528,041 (100 %)
Average mean PHQ-9 scores per patient per year, Mean (SD)	0.42 (1.32)	1.18 (2.39)
Average median PHQ-9 scores per patient per year, Mean (SD)	0.29 (1.26)	0.89 (2.38)
Patients with total number of scores in database, N (%)		
0 scores	334,596 (63.4 %)	228,557 (43.3 %)
1+ scores	193,445 (36.6 %)	299,484 (56.7 %)
2+ scores	98,715 (18.7 %)	230,440 (43.6 %)
3+ scores	74,934 (14.2 %)	202,259 (38.3 %)
4+ scores	62,314 (11.8 %)	180,658 (34.2 %)
5+ scores	53,218 (10.1 %)	161,434 (30.6 %)

multiple individual factors that can lead to variations in observed PHQ-9 scores (Ford et al., 2020; Ma et al., 2021). Therefore, some discrepancies in the agreement between the PHQ-9 and ePHQ-9 should be expected.

While certain biases may contribute to differences between the reported PHQ-9 and the ePHQ-9 scores, utilizing estimations in large EMR databases can also help to mitigate other biases inherent in small populations and sample bias between clinics that use the PHQ-9 consistently versus those that do not. Whether a patient receives a PHQ-9 assessment is influenced by multiple factors including clinicians' time limitations, lack of reimbursement from payers, severity of illness, language barriers or barriers to technology use (Kroenke, 2021). By applying a machine learning model to real-world data sources, the results of the current study show a substantial increase in the number of patients who automatically receive a PHQ-9 score based on the eligibility requirements of the model (i.e., data sufficiency in clinical notes) rather than access to the measurement tool.

Tracking symptom progression and severity in depression is vital to improve clinical decision-making and support research, enhance screening, evaluate treatments and improve outcomes (Gliklich et al., 2020). The lack of reported PHQ-9 scores in clinical practice inhibits the ability to measure treatment success and manage care for large numbers of patients with depression. This research contributes to addressing the measurement gap in mental health by developing and validating a model to generate ePHQ-9 scores from clinical notes using real-world data sources. The ability to fill in gaps in clinical encounters where no patient reported PHQ-9 score is documented increases the number of outcome measures available for clinical follow-up and provides a more comprehensive view of therapies over time. The result is more robust real-world data that can be used to assess patient journeys with greater fidelity and continuity to evaluate depression treatments for their effectiveness in the real-world. Future studies will assess the application of this novel machine learning model in direct comparison with observed patient reported scores to clinically relevant questions and examine the utility of incorporating large language models based on deep learning and more diverse clinical datasets.

CRediT authorship contribution statement

Pedro Alves: Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization, Writing – review & editing, Writing – original draft. **Carl D. Marci:** Investigation, Data curation, Conceptualization, Writing – review & editing. **Chandra J. Cohen-Stavi:** Investigation, Conceptualization, Writing – original draft. **Costas Boussios:** Supervision, Resources, Project administration, Methodology, Investigation, Formal analysis, Writing – review & editing, Writing – original draft. **Katelynn Murray Whelan:** Writing – review & editing, Writing – original draft.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Pedro Alves reports a relationship with OM1 Inc. that includes: employment and equity or stocks.

Carl Marci reports a relationship with OM1 Inc. that includes: employment and equity or stocks.

Chandra J Cohen-Stavi reports financial support was provided by OM1 Inc. Chandra J Cohen-Stavi reports a relationship with OM1 Inc. that includes: employment and equity or stocks.

Katelynn Murray Whelan reports administrative support, statistical analysis, and writing assistance were provided by OM1. Katelynn Murray Whelan reports a relationship with OM1 Inc. that includes: non-financial support.

Costas Boussios reports a relationship with OM1 Inc. that includes: employment and equity or stocks.

Acknowledgements

We would like to acknowledge the members of the OM1 Data Science Team, with special thanks to Jonathan Gerber, for their support throughout the development of this model.

References

Adekkanattu, P. *et al.* (2018). 'Ascertaining depression severity by extracting patient health Questionnaire-9 (PHQ-9) scores from clinical notes', *AMIA annual symposium proceedings.* AMIA Symposium, 2018, pp. 147–156. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6371338/.

Barkil-Oteo, H., 2013. 'Collaborative Care for Depression in Primary Care: How Psychiatry Could "Troubleshoot" Current Treatments and Practices', *Yale Journal of Biology and Medicine,* 86(2), pp. 139–146. Available at: PMID: 23766735; PMCID: PMC3670434.

Cheung, B.S.W., Murphy, J.K., Michalak, E.E., et al., 2023. Barriers and facilitators to technology-enhanced measurement based care for depression among Canadian clinicians and patients: results of an online survey. J. Affect. Disord. 320, 1–6. Available at: https://doi.org/10.1016/j.jad.2022.09.055.

Choudhary, S., Thomas, N., Ellenberger, J., Srinivasan, G., Cohen, R., 2022. A machine learning approach for detecting digital behavioral patterns of depression using nonintrusive smartphone data (complementary path to patient health Questionnaire-9 assessment): prospective observational study. JMIR Form. Res. 6 (5), e37736. https://formative.jmir.org/2022/5/e37736.

Coley, R., Boggs, J., Beck, A., Simon, G., 2021. Predicting outcomes of psychotherapy for depression with electronic health record data. J. Affect. Disord. Rep. 6 (1), e100198. Available at: https://doi.org/10.1016/j.jadr.2021.100198.

De Leeuw, J., Hornik, K., Mair, P., 2010. Isotone Optimization in R: Pool-Adjacent-Violators Algorithm (PAVA) and Active Set Methods. J. Stat. Softw. https://doi.org/10.18637/jss.v032.i05, 32(5), pp. 1–24. Available at.

Ford, J., Thomas, F., Byng, R., McCabe, R., 2020. Use of the Patient Health Questionnaire (PHQ-9) in Practice: Interactions between patients and physicians. Qual. Health Res. https://doi.org/10.1177/1049732320924625, 30(13):2146–2159. Available at.

Fraguas, R., Marci, C., Fava, M., et al., 2007. Autonomic Reactivity to Induced Emotion as a Potential Predictor of Response to Antidepressant Treatment. Psychiatry Res. https://doi.org/10.1016/j.psychres.2006.08.008, 151(1–2), pp. 169–172. Available at.

Gliklich, R., et al., 2020. Harmonized Outcome Measures for Use in Depression Patient Registries and Clinical Practice. Ann. Intern. Med. https://doi.org/10.7326/M19-3818, 172(12). Available at.

Goodwin, R., et al., 2022. Trends in U.S. Depression Prevalence From 2015 to 2020: The Widening Treatment Gap. Am. J. Prev. Med. 63 (5), 726–733. https://www.ajpmonline.org/article/S0749-3797(22)00333-6/fulltext.

Greenberg, P.E., et al., 2021. The Economic Burden of Adults with Major Depressive Disorder in the United States (2010 and 2018). PharmacoEconomics 39 (6), 653–665. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8097130/.

Hatton, C., et al., 2019. Predicting Persistent Depressive Symptoms in Older Adults: A Machine Learning Approach to Personalised Mental Healthcare. J. Affect. Disord. Rep. 246, 857–860. https://www.sciencedirect.com/science/article/abs/pii/S0165032718319931.

Hong, R.H., Murphy, J.K., Michalak, E.E., et al., 2021. Implementing Measurement-Based Care for Depression: Practical Solutions for Psychiatrists and Primary Care Physicians. Neuropsychiatr. Dis. Treat. 17 (1), 79–90. https://www.tandfonline.com/doi/full/10.2147/NDT.S283731.

Inoue, T., Tanaka, T., Nakagawa, S., et al., 2012. Utility and limitations of PHQ-9 in a clinic specializing in psychiatric care. BMC Psychiatry. https://doi.org/10.1186/1471-244X-12-73, 12 (73).

Kroenke, K., 2021. PHQ-9: global uptake of a depression scale. World Psychiatry Off. J. World Psychiat. Assoc. (WPA) 20 (1). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7801833/.

Kroenke, K., Spitzer, L., Williams, B., 2001. The PHQ-9: validity of a brief depression severity measure. J. Gen. Intern. Med. 16 (9). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1495268/.

Lee, B., Wang, Y., Carlson, S.A., et al., 2020. National, State-Level, and County-Level Prevalence Estimates of Adults Aged ≥18 Years Self-Reporting a Lifetime Diagnosis of Depression — United States', MMWR Morb Mortal Wkly Rep 2023, 72(1), pp. 644–650. https://www.cdc.gov/mmwr/volumes/72/wr/mm7224a1.htm?s_cid=mm7224a1_w#suggestedcitation.

Löwe, B., Unützer, J., Callahan, C.M., Perkins, A.J., Kroenke, K., 2004. Monitoring depression treatment outcomes with the patient health questionnaire-9. Med. Care 42 (12), 1194–1201. https://pubmed.ncbi.nlm.nih.gov/15550799/.

Ma, S., Kang, L., Guo, X., et al., 2021. Discrepancies Between Self-Rated Depression and Observed Depression Severity: The Effects of Personality and Dysfunctional Attitudes. Gen. Hosp. Psychiatry. https://doi.org/10.1016/j.genhosppsych.2020.11.016, 70 (May–June), pp. 25–30.

Place, S., Blanch-Hartigan, D., Smith, V., Erb, J., Marci, C.D., Ahern, D.K., 2020. Effect of a Mobile Monitoring System vs Usual Care on Depression Symptoms and Psychological Health: A Randomized Clinical Trial. JAMA Netw. Open 3 (1), e1919403. https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2758857.

Santomauro, D., et al., 2021. Global Prevalence and Burden of Depressive and Anxiety Disorders in 204 Countries and Territories in 2020 Due to the COVID-19 Pandemic. Lancet 398 (10312), 1700–1712. https://www.thelancet.com/action/showPdf?pii=S0140-6736%2821%2902143-7.

Schober, P., Mascha, E.J., Vetter, T.R., 2021. Statistics From A (Agreement) to Z (z Score): A Guide to Interpreting Common Measures of Association, Agreement, Diagnostic Accuracy, Effect Size, Heterogeneity, and Reliability in Medical Research. Anesth. Analg. https://doi.org/10.1213/ANE.0000000000005773, 133(6): pp. 1633–1641.

White, N., Parsons, R., Collins, G., et al., 2023. Evidence of Questionable Research Practices in Clinical Prediction Models. BMC Med. 21 (339). https://doi.org/10.1186/s12916-023-03048-6.

Xu, Z., et al., 2023. Using Machine Learning to Predict Antidepressant Treatment Outcome from Electronic Health Records. Psychiat. Res. Clin. Pract. 5 (4), 118–125. https://onlinelibrary.wiley.com/doi/10.1176/appi.prcp.20220015.

Zeier, Z., et al., 2018. Clinical Implementation of Pharmacogenetic Decision Support Tools for Antidepressant Drug Prescribing. Am. J. Psychiatry 175 (9), 813–916. https://ajp.psychiatryonline.org/doi/epdf/10.1176/appi.ajp.2018.17111282.

Zhou, L., et al., 2015. Identifying Patients with Depression Using Free-text Clinical Documents. Stud. Health Technol. Inform. 216, 629–633. https://pubmed.ncbi.nlm.nih.gov/26262127/.

Zhou, Z., Luo, D., Yang, B., Liu, Z., 2022. Machine Learning-Based Prediction Models for Depression Symptoms Among Chinese Healthcare Workers During the Early COVID-19 Outbreak in 2020: A Cross-Sectional Study. Front. Psych. 13 (1). https://doi.org/10.3389/fpsyt.2022.876995.

Zimmerman, M., McGlinchey, J., 2008. Why Don't Psychiatrists Use Scales to Measure Outcome When Treating Depressed Patients? J. Clin. Psychiatry. https://doi.org/10.4088/jcp.v69n1209, 69(12), pp. 1916–19.