Development of Machine Learning Model for Estimating PHQ-9 Scores Using Clinical Notes from Real-World Data Sources

Jonathan Gerber, PhD, Carl Marci, MD, Michelle Leavy, MPH, Pedro Alves, BS, Costas Boussios, PhD | OM1, Inc, Boston, MA, USA

Background and Rationale

The Patient Health Questionnaire-9 (PHQ-9) is a validated patient-reported measure for assessing depressive symptoms and symptom severity over the past two weeks. The PHQ-9 is widely used in both mental health and primary care settings because it is brief and easy to score (1).

Consistent capture of the PHQ-9 over time is important for monitoring changes in depressive symptoms, understanding response to treatment, and tracking outcomes such as remission and recurrence (2).

Yet, documentation of the PHQ-9 is inconsistent in real-world data sources such as electronic medical records (EMRs). This limits the potential role of these data sources for supporting large, heterogeneous research studies on depression treatment and outcomes. This effort aimed to apply machine learning methods to estimate PHQ-9 scores using routinely-recorded data from unstructured and semi-structured clinical notes

Objective

This effort aimed to apply machine learning methods to estimate PHQ-9 scores using routinely-recorded data from unstructured and semi-structured clinical notes.

Methods

A machine learning model was developed to generate estimated PHQ-9 (ePHQ-9) scores for specific clinical encounters using clinical notes from visits with mental health professionals. Data were drawn from the OM1 PremiOM Major Depressive Disorder (MDD) Dataset, a real-world data source containing data on over 490,000 MDD patients receiving treatment from mental health professionals across the United States.

Patients with both recorded PHQ-9 scores and clinical notes were identified and randomly assigned to a training cohort (32,802 patients contributing 96,891 encounters) to train an ensemble model when both notes and questionnaire data are available or a validation cohort (15,792 patients contributing 46,333 encounters). Notes were transformed via medical language processing, and the resulting features were reviewed and approved by a subject matter expert.

Methods (continued)

To assess model performance, the area under the receiver-operating-characteristic curve (AUC) was calculated using a binarized version of the outcome, and continuous ePHQ-9 scores were evaluated using Spearman R and Pearson R values. The approach used here has been used to develop validated estimated endpoints in other clinical areas, but this is the first application of this methodology in mental health (3-5).

Results

The model had an AUC of 0.81 when evaluating performance using the binarized version of the outcome with a cutoff of 9.5 in the validation cohort (i.e., the binary variable was high for PHQ-9 scores greater than 9.5 and low for all other scores). When evaluating performance using the continuous ePHQ-9 scores, the model had a Spearman R value of 0.62 and a Pearson R value of 0.61. Of note, model features included items similar to PHQ-9 questions (e.g., fatigue) as well as other concepts (e.g., medication usage, condition-specific scores).

When applied to the larger MDD Dataset, the model resulted in the generation of new ePHQ-9 scores for 2,215,662 (2.7x enrichment over 814,166 recorded PHQ-9's) encounters for 208,692 (1.2x enrichment over 174,897 patients with a PHQ-9) distinct patients.

Conclusions

- A machine learning model can estimate PHQ-9 scores using information routinely recorded in clinical notes from psychiatrists.
- At the individual patient level, use of the model could provide a more complete view of a patient's depressive symptom severity and response to treatment over time.
- At the population level, application of the model to RWD sources increases the number of patients and encounters available for research on depression treatment and outcomes.

Figure 1. Area Under the Receiver-Operating-Characteristic Curve (AUC)

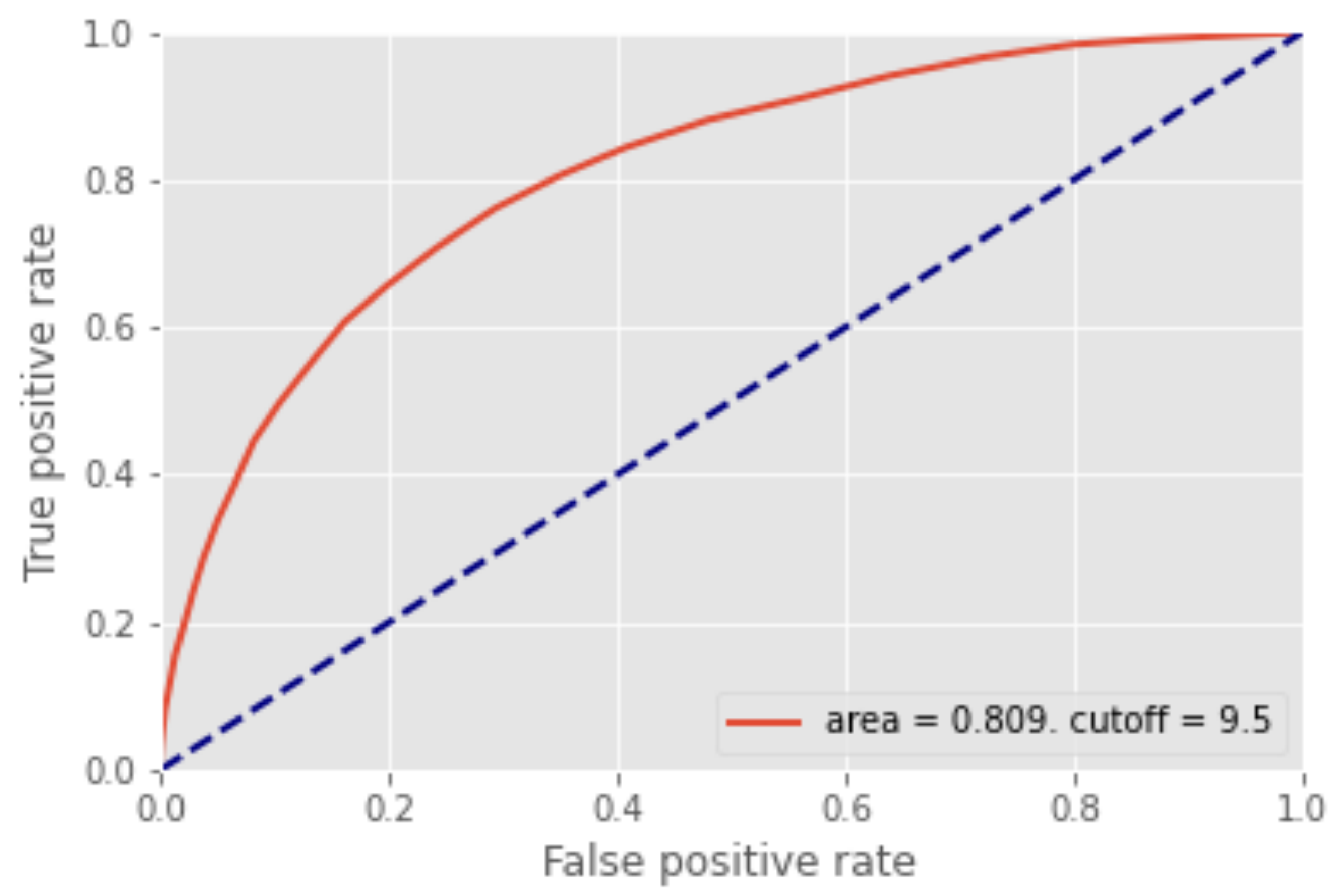
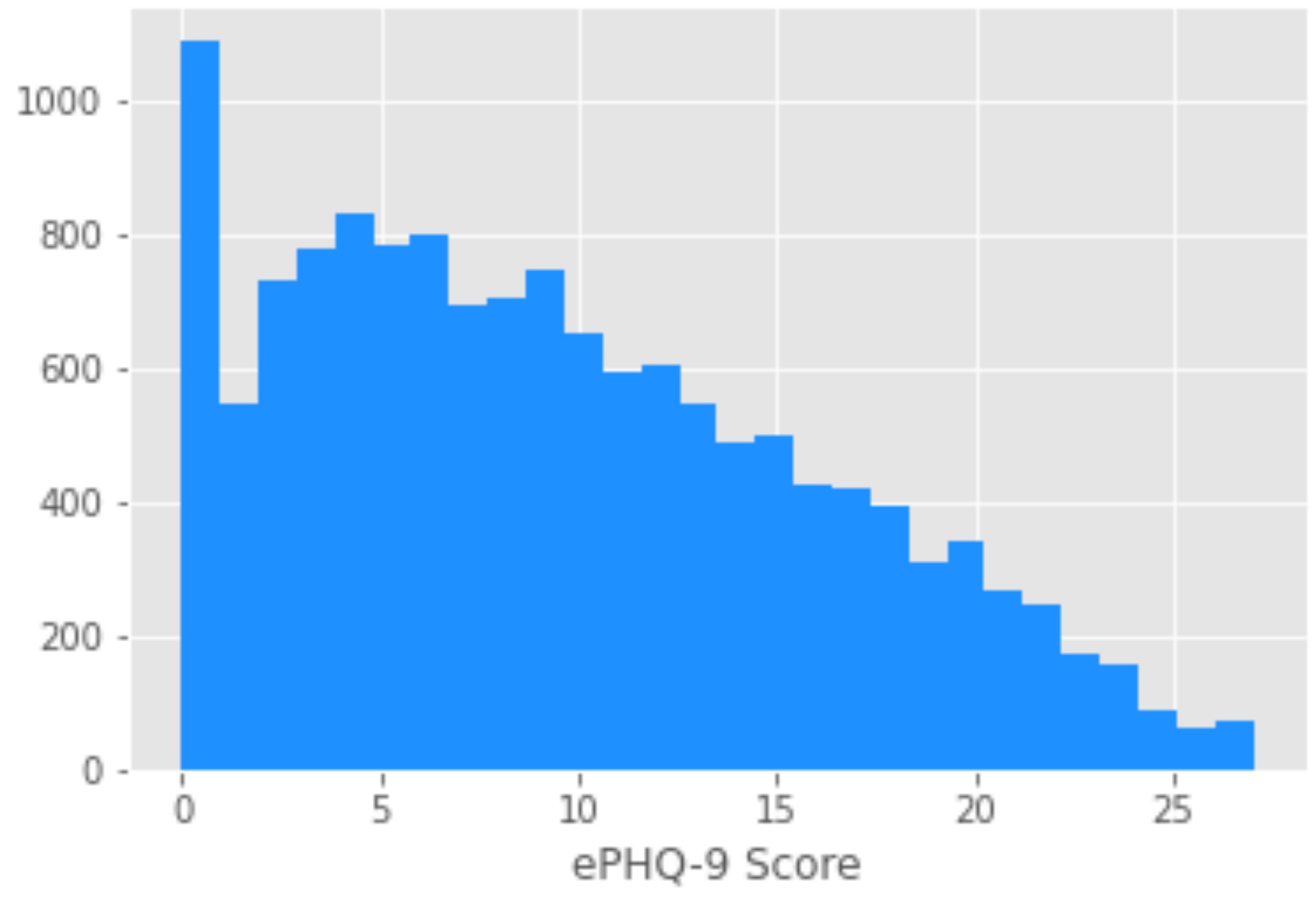


Figure 2. Distribution of ePHQ-9 Scores in Validation Cohort



Conference

Presented at 2023 ASCP Annual Meeting. May 30-June 3, 2023. Miami, FL, USA

References

1. Siu et al. Screening for Depression in Children and Adolescents: US Preventive Services Task Force Recommendation Statement. Pediatrics. 2016;137(3):e20154467.
2. Gliklich et al. Harmonized Outcome Measures for Use in Depression Patient Registries and Clinical Practice. Ann Intern Med. 2020;172(12):803-809.
3. Alves et al. Validation of a machine learning approach to estimate Systemic Lupus Erythematosus Disease Activity Index score categories and application in a real-world dataset. RMD Open. 2021;7(2):e001586.
4. Spencer et al. Validation of a machine learning approach to estimate Clinical Disease Activity Index Scores for rheumatoid arthritis. RMD Open. 2021;7(3)..
5. Alves et al. Validation of a machine learning approach to estimate expanded disability status scale scores for multiple sclerosis. Mult Scler J Exp Transl Clin. 2022;8(2):20552173221108635.