Augmenting Mental Health Decision-Making: a Temporal Symptom Trajectory Approach

Tanner O'Rourke*
two297@my.utexas.edu
University of Texas at Austin
Austin, TX, USA

Abstract

Timely identification of mental health condition trends remains challenging in routine mental-health care due to widening treatment gaps trends and fragmented electronic health records (EHR), making detection difficult in routine care. We construct a day-level synthetic data cohort from Synthea Coherent augmented with a PHQ-9 module, derive dense patient-day features (demographics, utilization, screening scores, and antidepressant coverage with PDC-90 and gap counters), and define three supervision targets with 30-day look-ahead windows: (i) non-adherence, (ii) relapse, and (iii) subtle deterioration. We train a multi-head XGBoost model on a hybrid design matrix that combines engineered numerics with a sparse bag-of-codes from an event stream, using patient-level splits, early stopping, and class-imbalance weighting, and feed results to the Gemini-3n-e2b-it model. Evaluation relies on AUROC, RMSC for medication gap days and confusion tables for a clinical threshold; Calibration reports to evaluate patient prediction summaries and model calibration are planned. Our approach is feasible and yields clinically significant characteristics for adherence, risk of relapse, and subtle deterioration. While usage of synthetic data limit external validity, the pipeline offers a blueprint for forked analyses and validation on real-world EHRs.

CCS Concepts

• **Applied computing** → **Health informatics**; *Health care information systems*; • **Computing methodologies** → **Supervised learning by classification**; *Ensemble methods.*

*Both authors contributed equally to this research.

Keywords

mental health, PHQ-9, Synthea Cohort, medication adhernece, deterioration detection, XG-Boost, EHR

1 Introduction

Mental health disorders such as depression remain among the leading causes of disability worldwide despite the burden of treatment gaps persisting [2]. Treatment plans are hampered by delayed recognition, fragmented care (both in practice and professional settings). Current tools that offer a shared language for symptom diagnosis, such as the Patient Health Questionnaire (PHQ-9, and variant PHQ-2), as well as other screening tests such as the GAD-7 (anxiety), AUDIT-C (Alcohol Use Disorders Identification Test), and DAST-10 (Drug Abuse Screening Test), offer frameworks for measuring mental health signals among real-world large cohorts and have led to improved care [5]. While these tests are widely supported and offer a validated, shared language for symptom change [1], they don't account for the trajectory of mental health changes over time such as situational triggers, complications that fall out of the primary care setting, and are bound to the existing utilities of traditional medical care. Advances in reasoning models have enabled the extraction of meaningful temporal patterns from high dimensional healthcare data. However, many existing approaches focus on prediction accuracy without prioritizing integration into existing workflows. Additionally, human-AI collaboration is proven to provide the highest diagnosis accuracy. Our framework relies on generating a transparent workflow that can be utilized by clinicians to provide more accurate diagnoses

Our work addresses this challenge by combining temporal modeling with fine tuned trend analysis of clinically salient features to

enable the generation of trajectory summaries that can be directly interpreted by clinicians. We utilize a high-volume, fully linked, synthetic depression screening dataset from Synthea Mitre Coherent synthetic EHR, with an augmented PHQ-9 module, to enable a reproducible setting. Our approach relies on sequence learning from four critical targets: (i) comorbidity blind spots—expected co-diagnoses that are absent given a patient's history; (ii) prescription mis-adherence via rolling gap-day-level proportion of days covered (PDC); (iii) 30-day relapse risk - a composite of emergency/inpatient utilization with mental-health codes, severe psychiatric events, and prolonged medication gaps; (iv) subtle deterioration signaling from forward-fed PHQ-9, PHQ-2, FAD-7, Audit-C, and DAST-10 test scores, utilization spikes, or antidepressant restarts. We train a XGBoost classifier to performs continuous, per-day surveillance of mental-health risk of these targets. These results are fed forward to the Gemini API using Chain-of-Thought reasoning to provide readable summaries of predictions made by the model.

2 Related Work

EHR psychiatric risk models. Large-scale psychiatry studies using clinical data demonstrate feasibility and value, including EMR-enabled investigations of treatment-resistant depression at scale [4]. For psychosis, EHR/NLP studies predict relapse with structured and unstructured signals, highlighting both promise and external-validity risks [5].

Depression screening trajectories. The PHQ-9 is widely validated and supported in primary care. Reviews support its use for tracking severity and change [2]. We utilize deterioration partly via change in PHQ-9 ≥ 5 within 30 days, aligning with minimal clinically important change thresholds [2].

Population context. National surveillance data show increasing depression prevalence from 2015–2020 and an expanding treatment shortfall, motivating tools that can surface at-risk patients for outreach [3].

Human–AI collaboration. Clinical studies show that calibrated, interpretable AI assistance can improve clinician accuracy. Inversely poorly designed support can mislead diagnosis, underscoring an importance for careful integration of reasoning models [1]. Our CoT summary layer is framed as a reasoning aid, not a label source.

3 Data

3.1 Data Source (Synthea Cohort)

We customized Synthea with a phq9.json module that emits the nine PHQ-9 items as discrete LOINC-coded Observations during ambulatory encounters. Module logic ensures PHQ-9 items, diagnoses, and medications occur in the same encounter, so Observations, Conditions, and Medications share patient/encounter IDs. This yields high-volume, fully linked synthetic screening data, suitable for reproducible research. We ingest the Coherent release (patients/encounters/conditions/medications/observations), normalize encounter classes (EMERGENCY, INPATIENT, OUTPATIENT, URGENTCARE, WELLNESS, OTHER), harmonize timestamps to day level, and standardize demographics. See References for information on reproducing our framework.

3.2 2. Screening, Lifestyle, Status

We use daily-last value extraction by patient-day with LOINC codes for screening tests and keyword fallbacks.

- PHQ-9 (44261-6), PHQ-2 (55758-7), GAD-7 (70274-6), AUDIT-C (75626-2), DAST-10 (82667-7).

We extract lifestyle attributes from LOINC codes and collapse to bins.

- Smoking status: (72166-2) $\rightarrow$ (never, former, current)
- alcohol use: (75626-2)$\rightarrow$ (none, moderate, heavy)
- pregnancy: (2106-3, 80384-1, 2112-1) $\rightarrow$ (0, 1)

3.2.1 Antidepressant Coverage & Adherence. Antidepressant coverage is constructed over a per-patient per-day coverage calendar, defined by start/stop times per-day coverage. Adherence $\in [0, 1]$ is derived as a 90-day trailing mean of days prescribed.

3.2.2 Utilization, Demographics. We define per-day counts by admission type bucket in PHQ-9 (44261-6), PHQ-2 (55758-7), GAD-7 (70274-6), AUDIT-C (75626-2), DAST-10 (82667-7) on a **7-day** rolling sum. Demographics are added to the per-day calendar as Age at day, and one-hot encoding for sex (Male and Female), and the top-6 most common races prevalent in the synthetic data.

3.2.3 Artifacts, Assumptions.

- Daily patient table (*daily*): One row per-patient per-day. The last 128 days is selected per patient to maintain dataset size for reproducibility.
- Event stream (*events*): Chronological events for all patients with event codes *dx*, *med*, *adm*, *obs* per event type, placed into per-day bag of codes (*ADM:<bucket>*)

4 Method

4.0.1 Labels. Labels use a 30-day look-ahead and fixed thresholds.

- Prescription Non-Adherence: PDC on 90-day rolling window less than 0.8.
- Relapse: Future occurrence of any relapse anchor: (i) same day admission **and** mental-health diagnosis, (ii) **severe** mental health condition diagnosis on given day, (iii) prolonged medication gap ($\geq$ 30 days)
- Deterioration: Within 30-day window: (i) PHQ-9 score spike ≥ 5, (ii) admission utilization spike ≥ 2, or (iii) medication start from uncovered to any covered day.

4.0.2 Training Table & Splits. We factorize for each (*patient, day*) a unique id *event_type|code* to provide code maps from a *patient_id* $\rightarrow$ type, code, key, freq. Data splits are done at a patient-level to avoid data leakages, with a stratified random sample into a **train 75%, val 10%, test 15%** split. The model received the daily event table, labels, and per-day code_ids which are numeric zero-filled and sorted by patient and date.

4.0.3 Model. We use XGBoost (gradient-boosted decision trees) on a sparse matrix made by stacking numeric features and code-token features side-by-side.

4.0.4 Training. We use 3 binary classification heads *y_relapse*, *y_det*, and *non_adherent_flag* evaluated with AUPRC (area under

the precision–recall curve), as well as a regression head (*y_gap_days*) to continuously predict days, with a squared error objective.

4.0.5 Evaluation. Per head we report AUROC, AUPRC (primary), and Precision@K (K = top 5% of days), and pick a max F1 threshold, and apply it to a confusion matrix at a chosen operating point; For regression, we use an RMSE on validation and test data and plot the residual probabilities. Outputs can be paired with a lightweight reasoning note (e.g., "PHQ-9 + 6/14d + ED last week + 18 uncovered days → high 30-day relapse risk") to support rapid review.

5 Results

5.0.1 Data Characteristics. The training data utilized was highly skewed towards non-depressive cases, which skews confusion matrix results. For example with deterioration, the model was able to identify 438 Pos-Pos cases correctly, while over 104,000 cases Neg-Neg correctly. This trend shows the importance of data collection

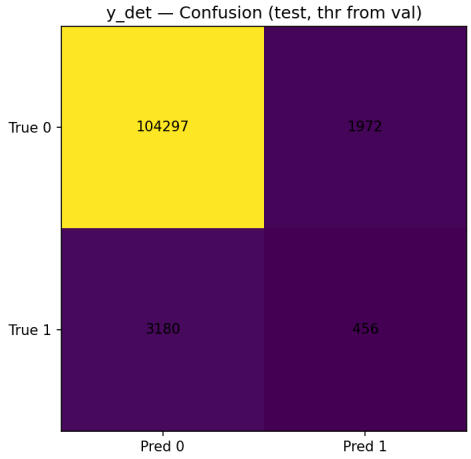

Figure 1: Confusion Matrix for MH Deterioration Head on test split data

and will be a further improvement of the study.

5.0.2 Head-specific analysis. This is further exemplified by ROC scores, which shows that the model's recall capabilities were particularly low and tended towards identifying depression cases negatively. This is likely due as well to the skew in the data, where many examples had a large prescription adherence, low deterioration, and few mental diagnoses.

6 Future Directions

Encode Relapse and Deterioration "Horizon" Days, Indicators for rise in PHQ9 symptoms, indicators for gap in adherence to medications

6.0.1 Data Synthesization. The data synthesis portion of this work stands out as a method for combining atomic patient EHR records into complex values of importance, however more learning is needed to be done by the researchers on proper dataset synthesization, especially in relation to class imbalances.

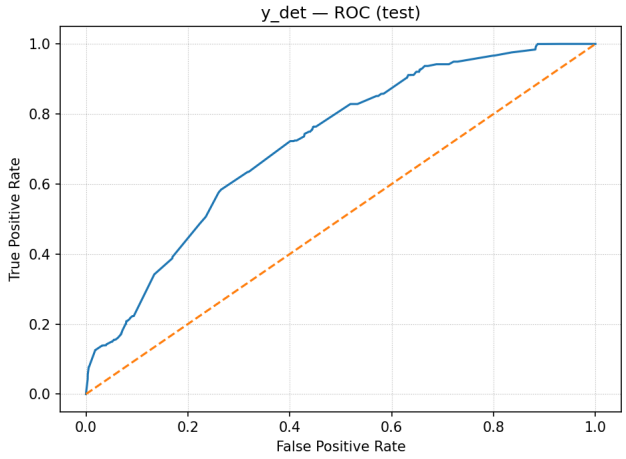

Figure 2: ROC score Curve for MH Deterioration on test split data

6.0.2 Limits on Synthetic Data. While this approach allows generation of high-volume, fully linked, synthetic depression screening datasets without accessing any protected health information, making it ideal for model development and reproducibility in mental health informatics research.

6.0.3 Prototyping. Finally, A prototype prototype and interface can be built to embedded EHR documents link directly to supporting timeline evidence, and capture clinician feedback for continuous learning. We can explore federated or split learning, and document a full reproducible training design that allows for a wider range of dataset ingestion.

6.0.4 Screener Gaps, LSTM. Screening tests are suitable baselines for synthesizing higher dimensional spaces and feature importance from root data, however further improvement can be made on the modeling choice. A better suited algorithm for this would be a Long-Term-Short-Term Memory RNN, which would describe time series in a more robust fashion.

References

A Running the Code

A.0.1 File Structure.

- /training-data: generated from Prepare-Data.ipynb, fed to XG-model.ipynb
- /out-data: Generated from Prepare-Data.ipynb, fed to Prepare-Data.ipynb
- /synthetic-data: Generated from Synthea Mitre Cohort, fed to Build-Features.ipynb
- /artifacts: plots generated from the last ran model
- XG-Model.ipynb: Create feature classification matrices, train XG Boost model
- Prepare-Data.ipynb: Transform classification data into ready to use training data, save to csv.
- Build-Features.ipynb: Build classification data from Synthea Mitre, save to csv Each respective file completes a respective

part of the model training steps. They should be run in the order Build-Features → Prepare-Data → XG-model. After each step, validate that the .csv files were uploaded to the respective folder.

A.1 Feature Codex

A.1.1 Labels.

- *non_adherent_flag*(`binary`)
- *y_relapse*(`binary, future 30 day`)
- *y_det*(`binary, future 30 day`)
- *y_gap_days*(`regression`)

A.1.2 Codebook. One binary column per code in your global vocabulary: code<id> These represent daily presence of:

- dx|<SNOMED> (conditions)
- med|<RX> (medications)
- adm|<BUCKET> (encounters)
- obs|<LOINC> (observations)

A.1.3 Adherence.

- ad_covered: 0/1 on that day for antidepressants
- pdc_90: rolling 90-day PDC; name reflects PDC_WINDOW_DAYS
- ad_gap_days: consecutive uncovered days up to today

A.1.4 Utilization & encounters.

- Encounter bucket one-hots per day: EMERGENCY, INPATIENT, OUTPATIENT, URGENTCARE, WELLNESS, OTHER
- util_7d: 7-day rolling sum of daily visits across buckets
- days_since_prev: gap to previous active day for that patient
- Screening scores totals: phq9, phq2, gad7, auditc, dast10

A.1.5 Lifestyle & pregnancy.

- Smoking 1-HE: smoke_never, smoke_former, smoke_current
- Alcohol one-hots: alcohol_none, alcohol_moderate, alcohol_heavy
- Pregnancy: pregnancy_pos → 0, 1

A.1.6 Demographics.

- age_years at encounter day
- sex_M, sex_F
- race_* - 1-HE top-6 races present in the data