EHR Foundation Models Improve Robustness in the Presence of Temporal Distribution Shift

Lin Lawrence Guo & Ethan Steinberg, et al.

Stanford University

October 20, 2023

Presented by Loh De Rong

Introduction

Purpose

- Temporal distribution shift negatively impacts the performance of clinical prediction models over time, and consequently its clinical utility (example: Epic Sepsis model).
- Pretraining EHR foundation models at scale is a useful approach for developing clinical prediction models that perform well in the presence of temporal distribution shift.

Intuition

- Compared ID and OOD performance of their EHR foundation model (CLMBR) to models trained on count-based representations
- Characterized the performance and robustness of different architecture choices for CLMBR, and how they scale with increasing quantity of pretraining data

Potential applications

- Clinical outcomes: hospital mortality, long length of stay, 30-day readmission, ICU admission
- Clinical diagnoses: coronary artery disease, SLE
- My work: autism, ADHD prediction

Background

Background knowledge required

- Medical concept embeddings
- Natural language processing
- CLMBR implementation

Methods: Overview

Objective: Demonstrate that global patterns embedded in CLMBR can generate more robust feature representations for downstream clinical tasks in the presence of temporal distribution shifts

Data source: STAnford medicine Research data Repository (STARR)

4 clinical outcomes (binary classification task)

1. Hospital mortality: patient death occurring during index admission
2. Long length of stay (*long LOS*): index admission of 7 or more days
3. Readmission in 30 days (*30-day readmission*): readmission to an inpatient unit within 30 days after discharge
4. Intensive care unit (*ICU*) admission: patient transfer to the ICU during the index admission

Methods: Patient Representations

Patient Representations

- Can be treated as a sequence of days that is ordered by time, $d_1 ... d_N$
- Each day consists a set of events represented by medical codes (e.g. diagnoses, lab tests, procedures and medications)

2 approaches of constructing patient representations

1. Count-based representations: binary features based on counts of both unique OMOP CDM concepts and derived elements recorded prior to the time of prediction.
2. Clinical Language Model-Based Representations (CLMBR): contatenate mean code embeddings with time delta features for each day, before feeding into the transformer or GRU
   - During adaptation, CLMBR weights were frozen, and a separate classification head was learned on the same patient representation for each clinical prediction task

Methods: Patient Representations

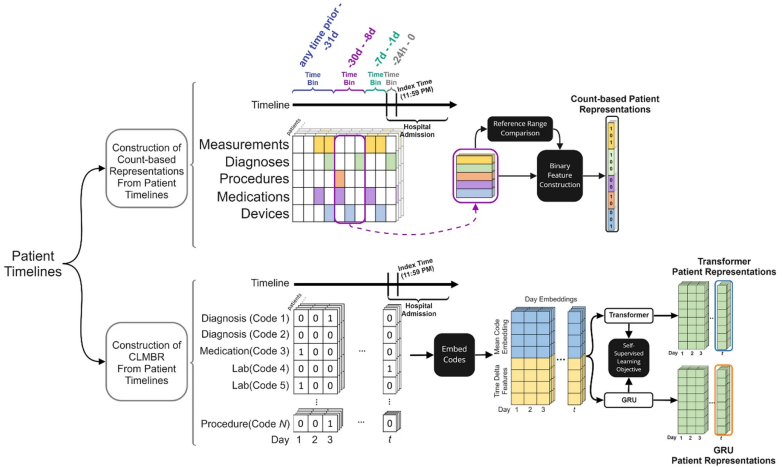

Figure: An overview of the two approaches of constructing patient representations used in this study.

Methods: Cohort Selection

2 types of cohorts

1. Pretraining: contained patients on which CLMBR was pretrained using the autoregressive objective
   - Number of patients: vary from 36K (42M coded events) to 1.8M (382M coded events)
   - Year range: vary depending on experimental setup
2. Task-specific: contained patients on which classification heads for clinical prediction tasks were trained and evaluated
   - Year range: from EHR inception (2009) to Aug 22, 2021
   - For patients with multiple admissions, one was randomly selected as the index admission for all tasks

Note: Patients in the task-specific cohort may overlap with the pretraining cohorts, however patients in the validation and test sets of the task-specific cohort were excluded from all pretraining cohorts in order to prevent data leakage.

Experimental Setup

Experiment 1: Compared ID and OOD performance of CLMBR-LR to count-LR

- Train: 2009-2012
- Validation: 2009-2012 (ID years)
- Test: 2009-2012 (ID years), 2013-2016 (OOD years), 2017-2021 (OOD years)

Experiment 2: Examined whether performance and robustness of CLMBR would improve with scale and whether the two CLMBR architectures scale differently

- Focused on scaling pretraining set size, from 36K to 1.8M patients

Experimental Setup

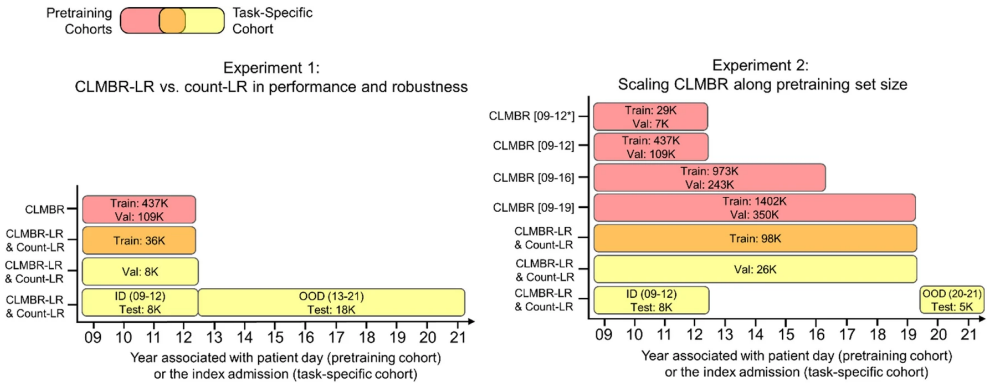

Figure: Year range for cohort selection and dataset size for CLMBR pretraining and task-specific models.

Implementation Details: Model Development

Count-based representations

- Additionally pruned features with less than 25 observations in the training set for each task separately, then pruned the same features from the validation and test sets
- Compared count-LR to oracle models (trained for each OOD year i.e. 2013-2021)

CLMBR

- Hyperparameter tuning (learning rate, dropout rate, batch size): grid search performed in Experiment 1 and subsequently used for each architecture in Experiment 2
- Compared CLMBR-LR to ETE with the same architecture (i.e. GRU and transformer).

After computing the representations, logistic regression with L2-regularization was trained for each clinical outcome in the task-specific training set of 2009-2012 in Experiment 1 and 2009-2019 in Experiment 2. Hyperparameter tuning was done on L2 regularization strength.

Implementation Details: Model Evaluation

Discrimination performance metrics

1. Area-under-the-receiver-operating-characteristic curve (AUROC)
2. Calibrated area-under the precision curve ($AUPRC_C$): computes the precision using a reference outcome prevalence, here set as the prevalence in the ID year group 2009-2012
3. Absolute calibration error (ACE): measure of calibration

Statistical analysis was subsequently performed using bootstrapping for each metric, and to compare models.

Experimental Results

Model degradation occurred in the OOD years (2013–2021) for long LOS and ICU admission prediction tasks, with larger degradations observed in 2017–2021.

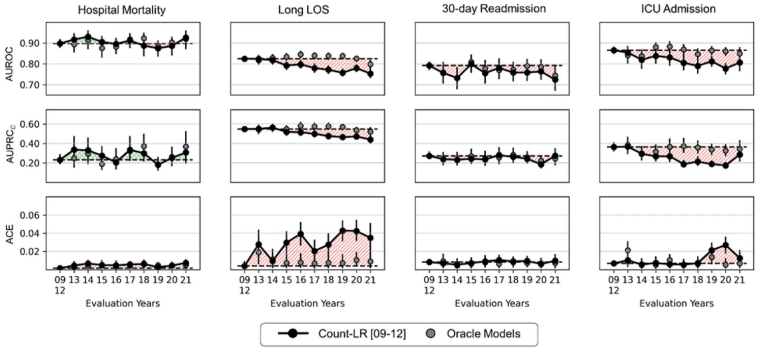

Figure: The impact of temporal distribution shift on the performance of count-LR.

Experimental Results

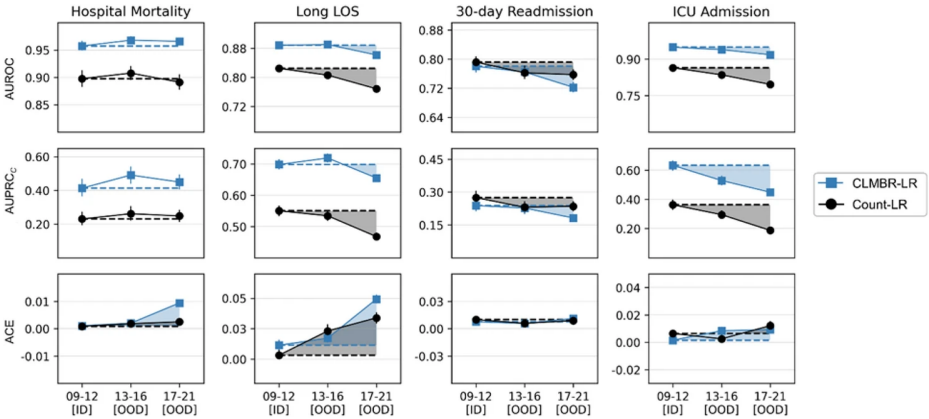

Figure: Performance of transformer-based CLMBR-LR and count-LR in the ID year group and their decay (shaded regions) in OOD year groups. GRU-based CLMBR results are qualitatively similar.

Experimental Results

Observations

1. CLMBR-LR outperformed count-LR in discrimination performance in both ID and OOD year groups across all tasks except for 30-day readmission.
2. CLMBR-LR displayed less degradation in AUROC and AUPRC$_C$ for long LOS and in AUROC in the 2017–2021 year group for ICU admission, whereas count-LR displayed less degradation in AUROC in the 2017–2021 year group for 30-day readmission.

Further Evaluations (see Supplementary Data)

1. CLMBR-LR performed as well as or better than its ETE counterpart in all tasks and metrics except for OOD calibration in 2017–2021 for long LOS and hospital mortality.
2. The performance of CLMBR models had strong positive correlations with the ID and OOD performance of their downstream logistic regression models in long LOS and ICU admission, and weak to strong positive correlation with hospital mortality.

Experimental Results

Generally, GRU-based CLMBR performed better with smaller pretraining set sizes whereas transformer based CLMBR exhibited better scaling of discrimination performance to pretraining set size, outperforming GRU-based CLMBR at larger pretraining set sizes.

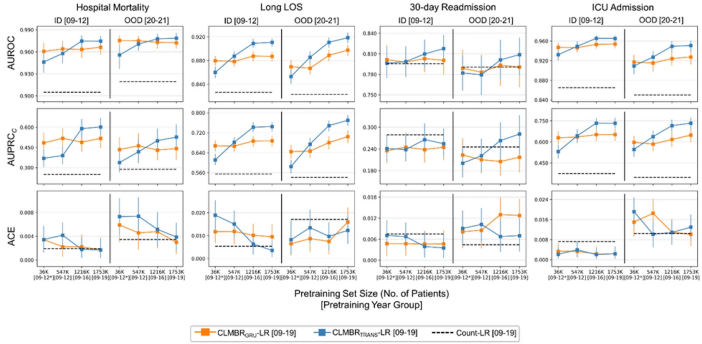

Figure: Scaling of GRU- and transformer-based CLMBR along pretraining set size.

Experimental Results

There were more correct re-classifications than incorrect re-classifications across risk thresholds and year groups for long LOS and ICU admission.

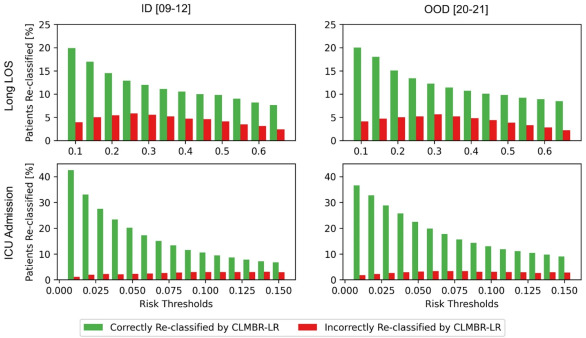

Figure: The proportion of patients correctly and incorrectly re-classified using CLMBR-LR instead of count-LR in long LOS and ICU admission.

Conclusion

Summary: Pretraining EHR foundation models at scale is a useful approach for developing clinical prediction models that perform well ID as well as OOD.

Things I liked:

- CLMBR learns a more holistic representation during pretraining that captures relationships between raw features (e.g. medical codes) - arguably more time-invariant, hence more robust to shifts
- The idea of having patient representation allows ML practitioners to focus on rapid adaptation of foundation models to downstream tasks (vs training ETE models from scratch)

Things I found lacking:

- Inadequate clinical explanations for observed trends
- Did not properly account for the worse calibration in CLMBR-LR

Recommendations:

- **Worth reading?** Yes. Good to know how we can evaluate our models under distribution shifts.
- **Worth implementing?** Yes. Can extend application to clinical diagnoses (binary classification, time-to-event prediction).