SCALED (Structured Comparison and Analysis of ML Experimental Design) Choices for Clinical Outcome Prediction

Kevin Shi, Ashley Phan, Hegler Tissot (Advisor)
Drexel University

2 Prediction Tasks
ED Disposition | ICU Readmission

BACKGROUND & MOTIVATION

Problem

- ML enables superhuman pattern recognition, but models can confidently make incorrect predictions
- Clinical prediction tasks impact patient care and hospital resource allocation
- How can we make clinically useful and realistic predictions?

Objectives

- Identify limitations of prior clinical ML approaches
- Preserve clinical interpretability through feature engineering
- Demonstrate the impact of threshold tuning on prediction balance
- Evaluate model performance using meaningful metrics

TASK FORMULATION

Feature Engineering

- Extracted features related to demographics, physiology, ICD codes, and hospital admission
- Used only prior-admission ICD codes to avoid data leakage
- Preserved missing values as NULL rather than imputing them

Datasets

Dataset	Instances	Features	Task
ED Disposition	268,651	13,233 (demographics, ICD codes, intime, vital signs)	Binary classification Class imbalance: 58.4% (Home) and 41.6% (Admitted)
ICU Readmission	79,003	1,404 (demographics, chart events, ICD codes, LOS, admission number)	Multi-label binary classification (readmission within 3/7/14/21/28/30 days) Class imbalance: 6.3%+ (3d) → 14.4%+ (30d)

Shared Modeling Framework

We trained separate XGBoost binary classifiers for each task:

1. Used StratifiedGroupKFold train/test splits to preserve class balance
2. Optimized the classification threshold using 5-fold CV, selecting the threshold that maximized validation F1 score in each fold
3. Re-trained the final model on the full training set using the mean optimal threshold across folds
4. Performed final evaluation on the held-out test set

IMPACT OF FEATURE SELECTION

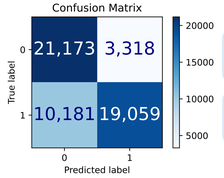
Fig. 1: Feretzakis et al. vs. Our Model (ED Disposition)

Baseline Model (Feretzakis et al.)		Our Model	
Patients 161,836	Features 28	Patients 214,920	Features 13,232

Reproduced prior methodology and preprocessing choices (excluding missing values) for comparison.

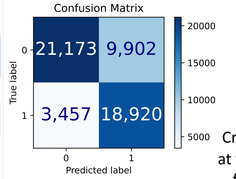
Modifies key ML design decisions (minimal exclusion and new features)

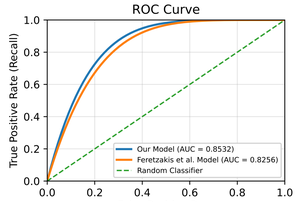
+53,084 patients | +13,204 features

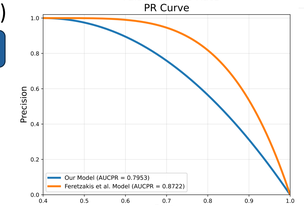


Same test set (n=53,731)

Same threshold





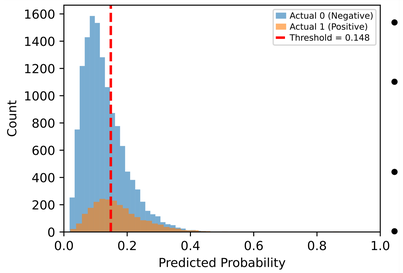


Cross dominance across ROC and PR-AUC curves at fixed threshold → the differences learned from feature representations, not threshold tuning

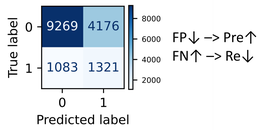
IMPACT OF THRESHOLD TUNING VS. UNDERSAMPLING

Fig. 2: Comparison of threshold tuning and undersampling predicted probability histograms and confusion matrices for the 30d label (ICU Readmission)

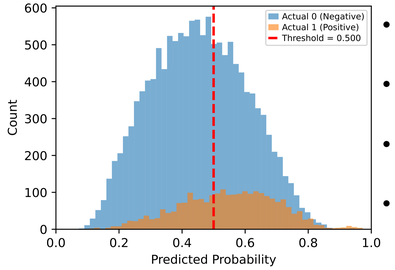
Threshold Tuning



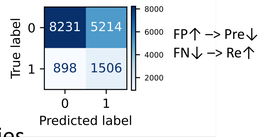
- Preserves original class distribution
- Adjusts the decision threshold directly to balance FP and FN
- Produces probabilities that better reflect real-world readmission risk → more clinically interpretable
- More adaptable to hospital capacity constraints



Undersampling



- Artificially balances the training data
- Shifts predictions toward the positive class
- Inflates predicted probabilities, making the default 0.5 threshold less realistic
- May overpredict ICU readmission risk under real-world class imbalance



RESULTS

Table 1: Model Performance by Feature Set (ED Disposition)

Model	Threshold	Precision	Recall	F1	ROC-AUC	PR-AUC
Baseline	0.357	0.852	0.652	0.739	0.826	0.872
Baseline 2.0	0.357	0.830	0.663	0.737	0.720	0.770
Our Model	0.357	0.656	0.845	0.739	0.853	0.795
Top 1 - ICD4019	0.328	0.583	0.859	0.695	0.776	0.699
Top 2 - ICD2500	0.363	0.609	0.855	0.711	0.754	0.702

Table 2: Lin et al. Reproduction (30d ICU Readmission)

Model	Threshold	Precision	Recall	F1	ROC-AUC
Reproduction (Undersampling)	0.5	0.379	0.712	0.495	0.791
Threshold Tuning	0.222	0.439	0.606	0.509	0.796

Table 3: Threshold Tuning Results (Multi-label ICU Readmission)

Label	Threshold	Precision	Recall	F1	ROC-AUC	PR-AUC
3d	0.104	0.179	0.237	0.204	0.650	0.170
7d	0.111	0.176	0.406	0.245	0.656	0.207
14d	0.122	0.193	0.534	0.283	0.659	0.241
21d	0.135	0.217	0.550	0.311	0.662	0.261
28d	0.149	0.239	0.535	0.330	0.666	0.273
30d	0.148	0.240	0.550	0.334	0.668	0.279

Table 4: Undersampling Results (Multi-label ICU Readmission)

Label	Threshold	Precision	Recall	F1	ROC-AUC	PR-AUC
3d	0.5	0.101	0.588	0.173	0.640	0.155
7d	0.5	0.147	0.579	0.234	0.655	0.200
14d	0.5	0.185	0.600	0.282	0.655	0.231
21d	0.5	0.207	0.601	0.307	0.655	0.250
28d	0.5	0.221	0.618	0.326	0.660	0.267
30d	0.5	0.224	0.626	0.330	0.665	0.274

CONCLUSIONS

1. AI/ML should support, not replace, clinical decision-making
2. Clinically useful and realistic ML models must balance FP and FN to support patient care and hospital resource allocation:
   a. Feature engineering influences the utility of clinical ML models
   b. Threshold tuning adapts operating points to varying hospital needs

MAIN REFERENCES

Cross, J.L. et al. *Bias in medical AI: Implications for clinical decision-making*, 2024.
Feretzakis, G. et al. *Machine Learning in Medical Triage: A Predictive Model for Emergency Department Disposition*, 2024.
Lin, Y.W. et al. *Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory*, 2019.
Riley, R.D. et al. *Stability of clinical prediction models developed using statistical or machine learning methods*, 2023.