Comparison of deep learning and conventional methods for disease onset prediction

Luis H. John[1], Chungsoo Kim[2], Jan A. Kors[1], Junhyuk Chang[3], Hannah Morgan-Cooper[4], Priya Desai[4], Chao Pang[5], Peter R. Rijnbeek[1], Jenna M. Reps[1,6], Egill A. Fridgeirsson[1]

[1]Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands; [2]Section of Cardiovascular Medicine, Department of Internal Medicine, Yale School of Medicine, New Haven, CT, United States; [3]Department of Biomedical Informatics, Ajou University Graduate School of Medicine, Suwon, Republic of Korea; [4]Stanford School of Medicine and Stanford Health Care, Palo Alto, CA, United States; [5]Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, NY, United States; [6]Janssen Research and Development, Titusville, NJ, United States

Background: Identifying individuals at high risk of disease at an early stage allows for improved care and risk-factor targeted intervention. Conventional approaches such as logistic regression and gradient boosting (XGBoost) have long served as reliable tools for predictive modeling in the clinical domain. However, the continuous advancement of deep learning methods, such as ResNet and Transformer, offers the promise of improved prediction accuracy and the ability to extract intricate patterns from complex clinical data. This study compares these conventional and deep learning methods to predict dementia in persons aged 55 – 84, bipolar disorder in patients newly diagnosed with major depressive disorder, and lung cancer in patients aged 45 – 65. We use observational data from administrative claims and electronic health records mapped to the OMOP CDM and follow the standardized OHDSI patient-level prediction approach for onset prediction in Figure 1.

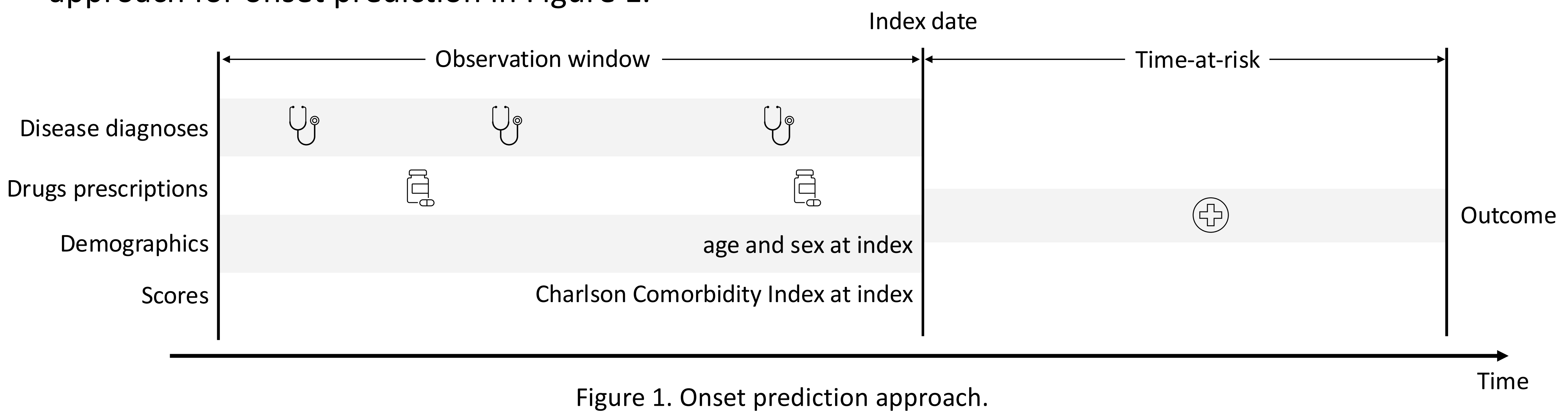
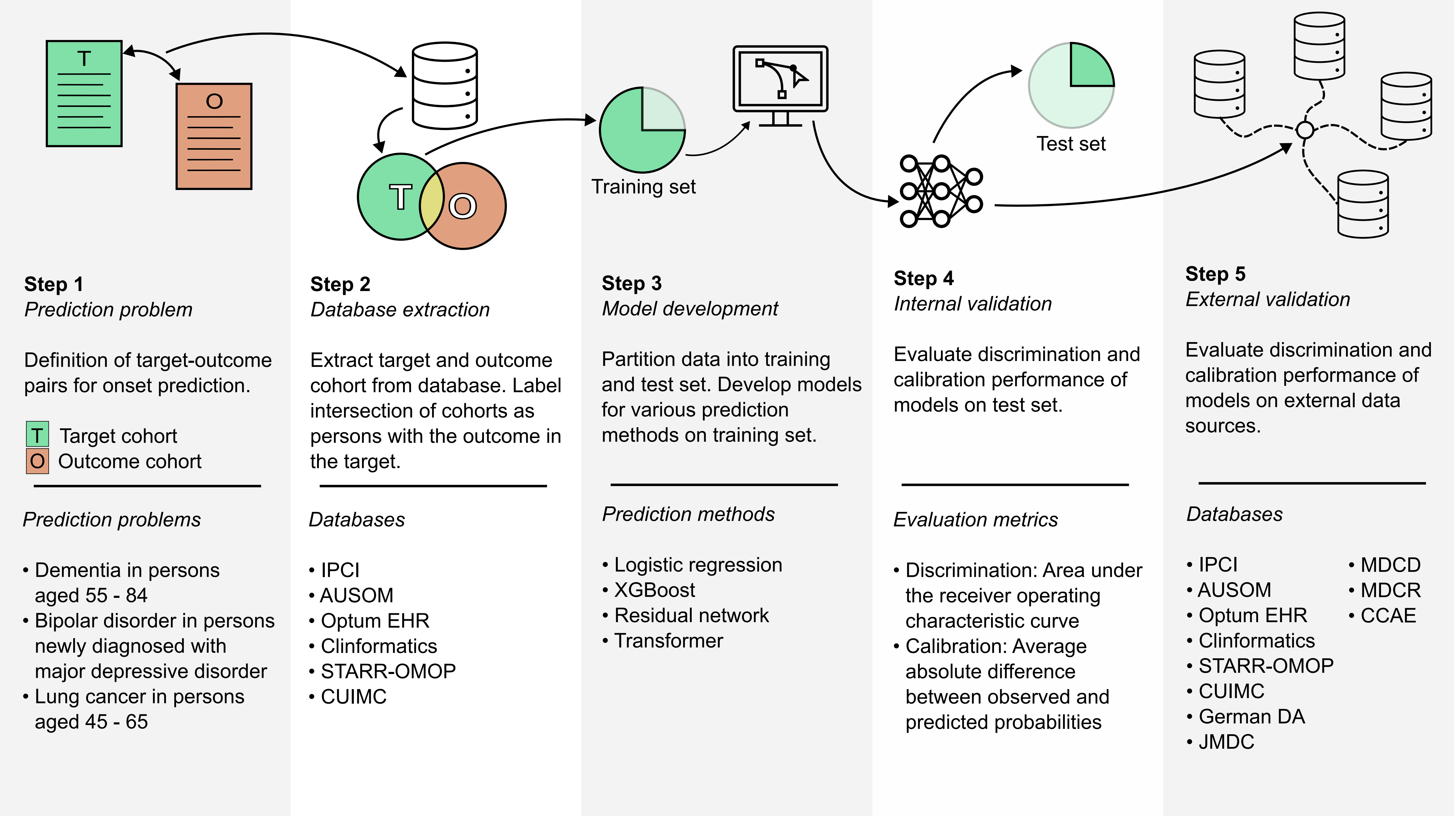


Figure 1. Onset prediction approach.

Methods: A study overview is presented in Figure 2. We evaluate internal and external validation performance using AUROC for discrimination and E_{avg} for calibration. Friedman's test is used to detect ranking differences of the different prediction methods. If the null hypothesis for no difference in ranks between the methods is rejected, we proceed with a post-hoc test to examine all pairwise differences, controlling for multiplicity. The results are plotted in a critical difference (CD) diagram of the Nemenyi test, which shows the mean ranks of each prediction method.



Step 1
Prediction problem

Definition of target-outcome pairs for onset prediction.

T Target cohort
O Outcome cohort

Prediction problems

- Dementia in persons aged 55 - 84
- Bipolar disorder in persons newly diagnosed with major depressive disorder
- Lung cancer in persons aged 45 - 65

Step 2
Database extraction

Extract target and outcome cohort from database. Label intersection of cohorts as persons with the outcome in the target.

Databases

- IPCI
- AUSOM
- Optum EHR
- Clinformatics
- STARR-OMOP
- CUIMC

Step 3
Model development

Partition data into training and test set. Develop models for various prediction methods on training set.

Prediction methods

- Logistic regression
- XGBoost
- Residual network
- Transformer

Step 4
Internal validation

Evaluate discrimination and calibration performance of models on test set.

Evaluation metrics

- Discrimination: Area under the receiver operating characteristic curve
- Calibration: Average absolute difference between observed and predicted probabilities

Step 5
External validation

Evaluate discrimination and calibration performance of models on external data sources.

Databases

- IPCI
- AUSOM
- Optum EHR
- Clinformatics
- STARR-OMOP
- CUIMC
- German DA
- JMDC
- MDCD
- MDCR
- CCAE

Figure 2. Study overview.

Contact: l.john@erasmusmc.nl

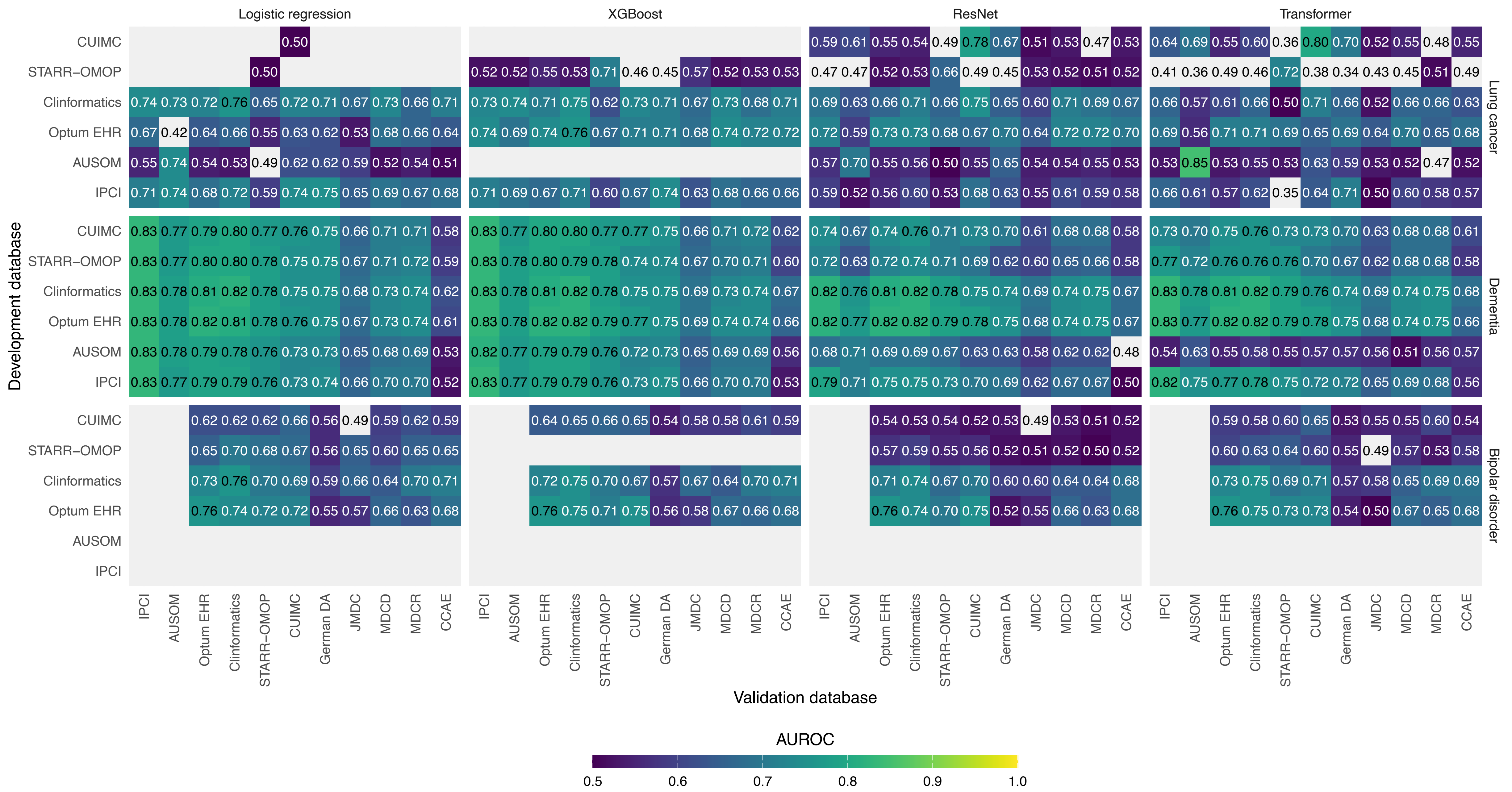


Figure 3. Internal and external discrimination performance (AUROC) across prediction methods and prediction problems.

Discussion: Discrimination performance across databases, prediction methods, and prediction problems is presented in Figure 3. Using these measures, the CD diagram in Figure 4A reveals that conventional methods outperform deep learning methods. However, assessing only internal validation performance, no significant difference between methods is found and no post-hoc test is performed. This is confirmed by learning curve analysis in Figure 5, which shows that performance of conventional and deep learning methods converges if enough data is available. Conventional models transport better (Figure 4B) and rank better on small data (Figure 4C). Small data also causes poor calibration in ResNet.

Our finding highlights the current limitations of deep learning methods when applied to observational healthcare data. These methods are more complex and require more data to train, but do not show better performance than conventional methods. However, the type of data we use, flattened tabular data, likely does not exploit the full capabilities of deep learning methods. Future work should focus on techniques that utilize the temporal nature of observational data to fully take advantage of the complex nature and pattern recognition capabilities of deep learning.

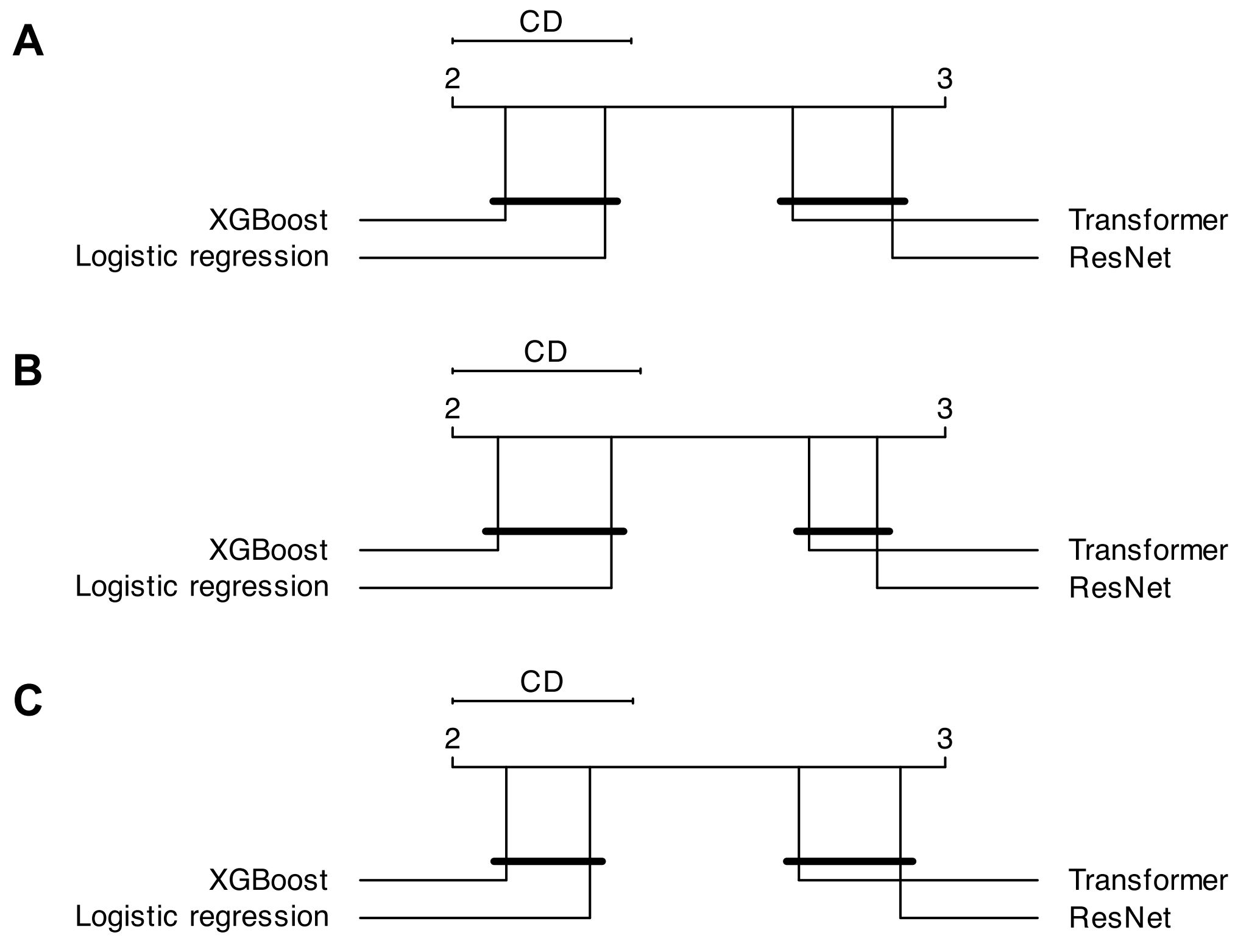


Figure 4. Ranking of prediction method based on AUROC for (A) internal and external validation, (B) external validation, (C) models developed on small data.

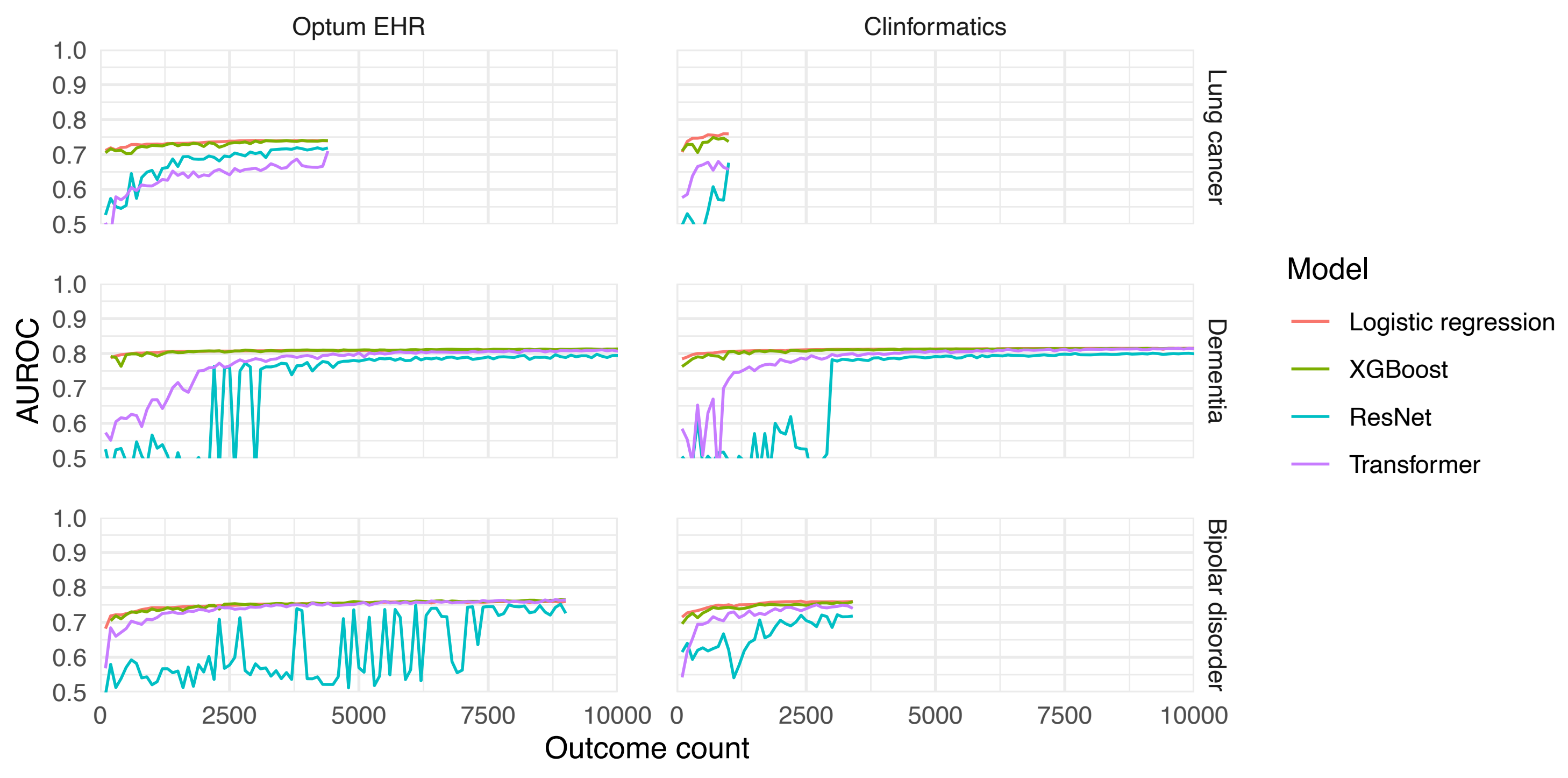


Figure 5. AUROC performance on the test set for increasingly larger subsets of the training set