Supervised Machine Learning to Predict HIV Outcomes Using Electronic Health Record and Insurance Claims Data

John Semerdjian, Konstantinos Lykopoulos, Andrew Maas, Morgan Harrell, Julie Priest, Pedro Eitz-Ferrer, Andrew Zolopa, Conor Wyand

Background

We aimed to develop a machine learning model that could predict key HIV outcomes, such as viral suppression, retention, and adherence to antiretroviral therapy (ART), using electronic health record (EHR) data. This study was part of a larger project that integrated natural language processing (NLP) and machine learning techniques to analyze electronic health records (EHRs) and claims data to predict HIV outcomes.

Methods

1. **Inclusion Criteria**: Any HIV ICD code was included in the dataset. The criteria were designed to capture patients with confirmed HIV diagnoses.

2. **Data Preparation**:
   - **Viral Load Suppression**: This metric was used to assess whether patients were virally suppressed, i.e., their viral load was below the threshold level.
   - **AUROC**: The Area Under the Receiver Operating Characteristic Curve was calculated to evaluate the model's performance.

3. **Parameters**:
   - Inclusion criteria: Any HIV ICD code.
   - Medication Adherence: This metric measured how well patients were adhering to their ART medication regimen.

4. **Models**:
   - **Word Embedding Details**: Created 2.5%, 50%, and 97.5% quantiles for metrics via 100 bootstrapped models tested against validation set.
   - **Predictive Model Details**: Processed 80/10/10 data split into training, test, and validation sets.

5. **Visualization**: Used a neural network visualization tool to interpret the model's predictions.

6. **Metrics**:
   - **Viral Load Suppression**: AUROC 1,126, AUC 0.83 (0.82–0.85), F1 score 0.78 (0.75–0.80). Retained in Care: AUROC 2,062, AUC 0.83 (0.82–0.85), F1 score 0.71 (0.70–0.72).
   - **ART Adherence**: AUROC 15,552, AUC 0.83 (0.82–0.83), F1 score 0.66 (0.65–0.66).

7. **References**:
   - Distributed representation of words and phrases and their compositionality, Mikolov et al., NIPS 26, 3111–3119, 2013.