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

WEPEC32

John Semerdjian,[1] Konstantinos Lykopoulos,[2] Andrew Maas,[1] Morgan Harrell,[1] Julie Priest,[3] Pedro Eitz-Ferrer,[3] Andrew Zolopa,[3] Conor Wyand[4]

[1]Roam Analytics; [2]ViiV Healthcare UK; [3]ViiV Healthcare US; [4]Optum Life Sciences

Corresponding Author: Dr John Semerdjian e: john@roaminsight.com

Introduction

The HIV care continuum in the US is still short of achieving the UNAIDS 90-90-90 target. Learning predictors for HIV outcomes may facilitate interventions that improve patient care, and achieve UNAIDS goals. We aimed to learn predictors for viral suppression, retention, linkage to care and antiretroviral therapy (ART) adherence. We discuss the application of machine learning methods to predict these HIV patient outcomes and opportunities for HIV care improvement based on HIV outcome predictors.

Methods

We selected HIV patient cohorts from a set of electronic health records (EHR) data and administrative claims data from Optum's de-identified Integrated Claims-EHR dataset (2007-2016). We constructed the following HIV Care Continuum outcome measures:

- **Viral Load Suppression:** <200 copies/mL, <2.3 log10(copies)/mL, or with undetectable HIV (qualitative) on most recent viral load (VL) test
  - Inclusion criteria: Any HIV ICD-9/10 code for HIV (symptomatic, asymptomatic, counseling), ≥1 VL test at any point during data collection period, and whose most recent VL test was in 2015 or 2016.
- **Retention / Engagement:** ≥2 CD4 or VL tests (or 1≥ for Engagement) in CY 2016 at least ≥3 months apart.
  - Inclusion criteria: Any HIV ICD-9/10 code for HIV (symptomatic, asymptomatic, counseling), ≥1 VL test at any point during data collection period, and had at least 2 EHR visits in 2015.
- **Linkage to Care:** ≥1 VL or CD4 test within 1 month of HIV diagnosis.
  - Inclusion criteria: HIV+ lab result in 2014-2016, and had no CD4 or VL lab tests before their earliest HIV+ diagnostic lab test.
- **Medication Adherence:** Proportion of days covered (PDC) where the adherence/non-adherence threshold was 80% (i.e. a patient with ≥ 80% PDC was considered Adherent).
  - Inclusion criteria: Any HIV ICD-9/10 code for HIV (symptomatic, asymptomatic, counseling), ≥1 VL test between 2014-2016, ≥1 non-Truvada ART prescription, and had at least 80% insurance coverage in calendar year prior to the outcome measure

Predictive Model Details

- **Models:** Logistic Regression with L2 Regularization using Python package Scikit-Learn[1]
- **Features:** All demographic, clinical (comorbidity, medications, provider specialty), encounter, and clinical text data from the 12 months prior to the outcome; converted into binarized values that represent presence/absence of feature (Figure 1)
- **Data preparation:** 80/10/10 data split into training, test, and validation sets
- **Parameters:** Regularization strength determined via 5-fold cross-validation
- **Metrics:** Created 2.5%, 50%, and 97.5% quantiles for metrics via 100 bootstrapped models tested against validation set
  - **AUROC:** Probability that the classifier will assign a higher score to a randomly chosen positive example than to a randomly chosen negative example
  - **F1:** Measures predictive performance that takes into account precision (i.e. positive predictive value) and recall (i.e. sensitivity)

Word Embedding Model Details

- **Models:** Word2Vec[2] and GloVe[3]
- **Features:** All clinical phrases for each patient patient record from the past 12 months (clinical term sequence unknown)
- **Data preparation:** Shuffled clinical phrases used for word2vec since clinical phrase sequence is unknown; patient-clinical term co-occurrence matrix used for GloVe
- **Parameters:** Skip-gram model for word2vec; 100-dimensional output vector for both algorithms
- **Visualization:** UMAP algorithm[4] converted 100-dimensional vectors into 3-dimensional vectors (Figure 2)

Figure 1. Example Model Vector Representation

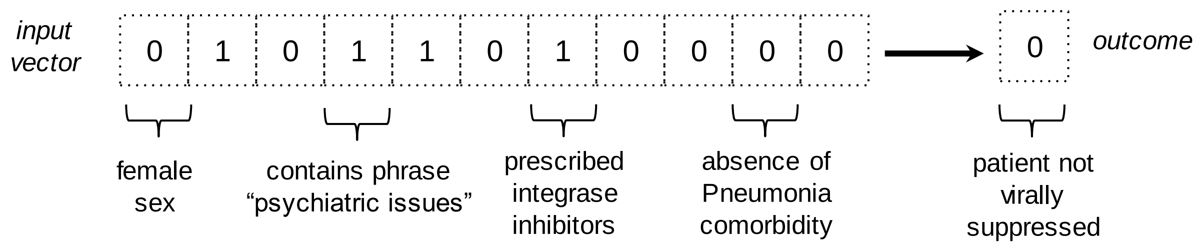

Figure 2. Example Text Embedding for target phrase "feel depression"

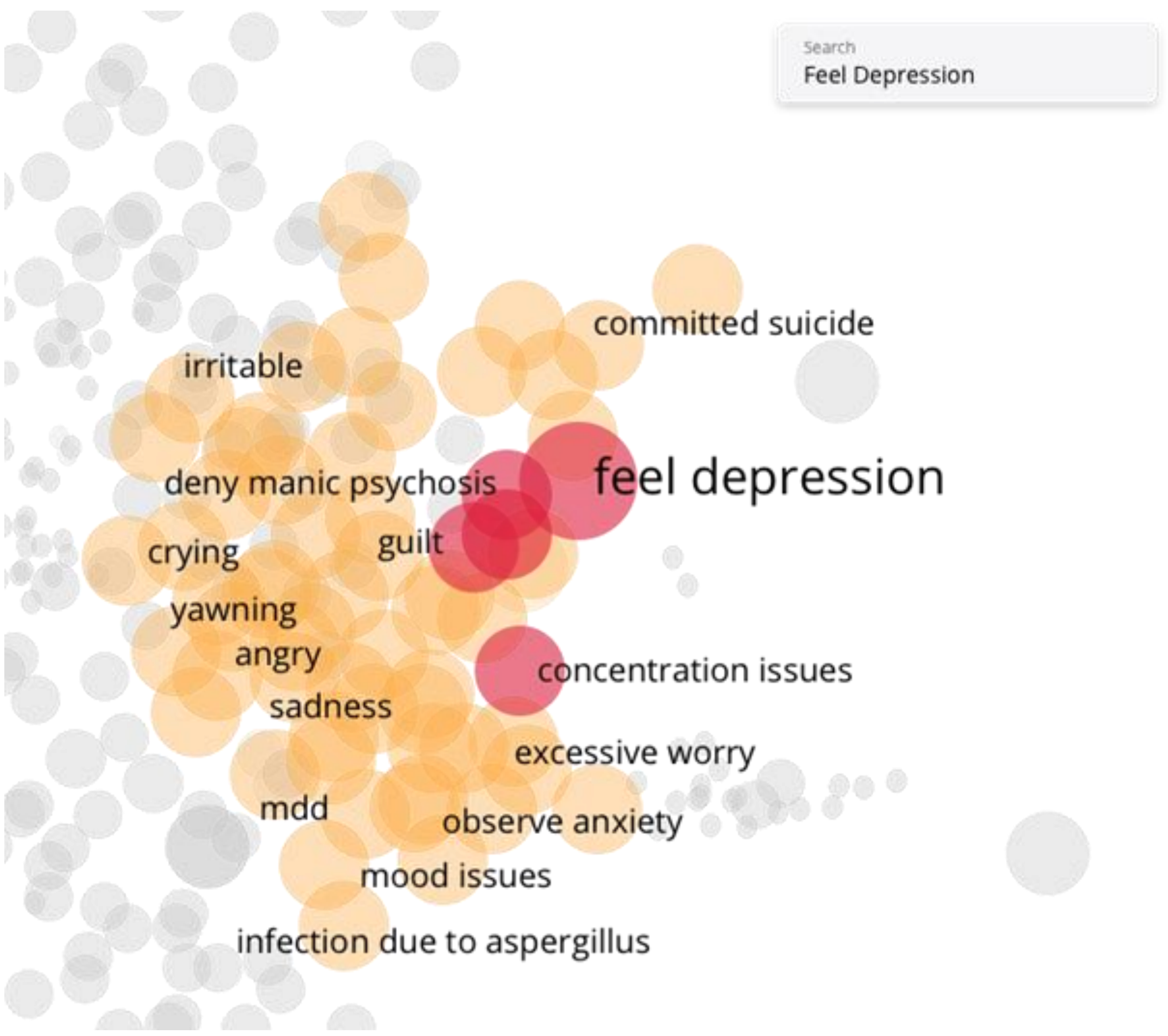

Table 1. Model Results

Models	Features	AUROC (95% CI)	F1* (95% CI)	No. features	Patients
Linkage to Care 30 days	Demographic-only	0.59 (0.55-0.62)	0.90 (0.90-0.90)	33	2,062
	All features	**0.63 (0.57-0.66)**	**0.90 (0.89-0.91)**	241	
Retained in Care [Engaged in Care]	Demographic-only	0.65 (0.64-0.65) [0.63 (0.63-0.64)]	0.66 (0.65-0.67) [0.63 (0.62-0.64)]	36 (36)	21,631 [22,649]
	Clinical text-only	0.73 (0.72-0.74) [0.73 (0.72-0.74)]	0.69 (0.68-0.70) [0.68 (0.67-0.69)]	1,230 (1,242)	
	All features	**0.75 (0.74-0.75) [0.75 (0.74-0.75)]**	**0.70 (0.69-0.71) [0.69 (0.68-0.70)]**	1,572 (1,572)	
Viral Load Suppression	Demographic-only	0.75 (0.74-0.75)	0.78 (0.77-0.78)	36	15,552
	Clinical text-only	0.83 (0.82-0.83)	0.83 (0.82-0.83)	814	
	All features	**0.83 (0.83-0.84)**	**0.83 (0.82-0.83)**	1,126	
ART Adherence Claims-only	Demographic-only	0.57 (0.57-0.58)	0.80 (0.80-0.80)	17	13,456
	All features	**0.66 (0.65-0.66)**	**0.80 (0.79-0.80)**	273	

*F1 score is a measure of predictive performance that takes into account precision (i.e. positive predictive value) and recall (i.e. sensitivity)

Discussion

To our knowledge, this poster presents the first attempt at modeling all HIV Care Continuum outcome measures using ML techniques. We assessed the relative difference in predictive performance across an array of common patient features, and discovered the predictive value of clinical text features, especially phrases associated with mental and behavioral health issues. Our ML models and word vector representations consider thousands of features simultaneously. These new capabilities will help healthcare providers, payors, and pharmaceutical companies generate new hypotheses regarding HIV care.

Conclusion

This study achieved moderately accurate prediction for HIV outcomes by applying natural language processing and machine learning to EHR and claims data. We suggest that reliable prediction for HIV outcomes may be held in the unstructured patient notes, and can be derived from natural language processing and machine learning.

References

1. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
2. Distributed representations of words and phrases and their compositionality, Mikolov et al., NIPS 26, 3111-3119, 2013.
3. GloVe: Global Vectors for Word Representation, Pennington et al., EMNLP, 1532-1543, 2014.
4. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, McInnes, L. Healy, J. ArXiv e-prints, 2018.

22nd International AIDS Conference; July 23-27, 2018; Amsterdam, the Netherlands