Evaluation of Study Execution using Large-Scale Analytics: A Machine Learning Approach to Assess Pre-Exposure Prophylaxis (PrEP) Utilization in the Real-World

Nag Mani[1], Xiwen Huang[1], Li Tao[1], Hu Li[1]
[1] Gilead Sciences

Background

Real-world evidence studies aim to generate reliable evidence using individual level observational data.[1] Understanding the variations and commonalities across data sources will contribute to the development of accurate and reliable evidence, ultimately driving improvements in fit-for-purpose interventions in the real world. Increasingly, studies have utilized a large-scale analysis (LSA) platform and machine learning methodology to address challenges when analyzing real-world data.

In this study, we explore the methodology of leveraging the LSA platform to perform a machine-learning assessment across three distinct data sources, aiming to demonstrate the strengths and limitations of such a multi-dataset analysis. Our work focuses on identifying populations who would benefit from HIV pre-exposure prophylaxis (PrEP) in real-world settings in the United States (US). Despite the proven effectiveness of PrEP, its uptake remains low among key populations vulnerable for HIV infection.[2,3] In 2020, approximately 25% of the 1.2 million people who would benefit from PrEP (PWBP) were prescribed PrEP.[3] This disparity motivates our study as we aim to provide a comprehensive characterization across data sources and compare the predictors for PrEP use in each data source.

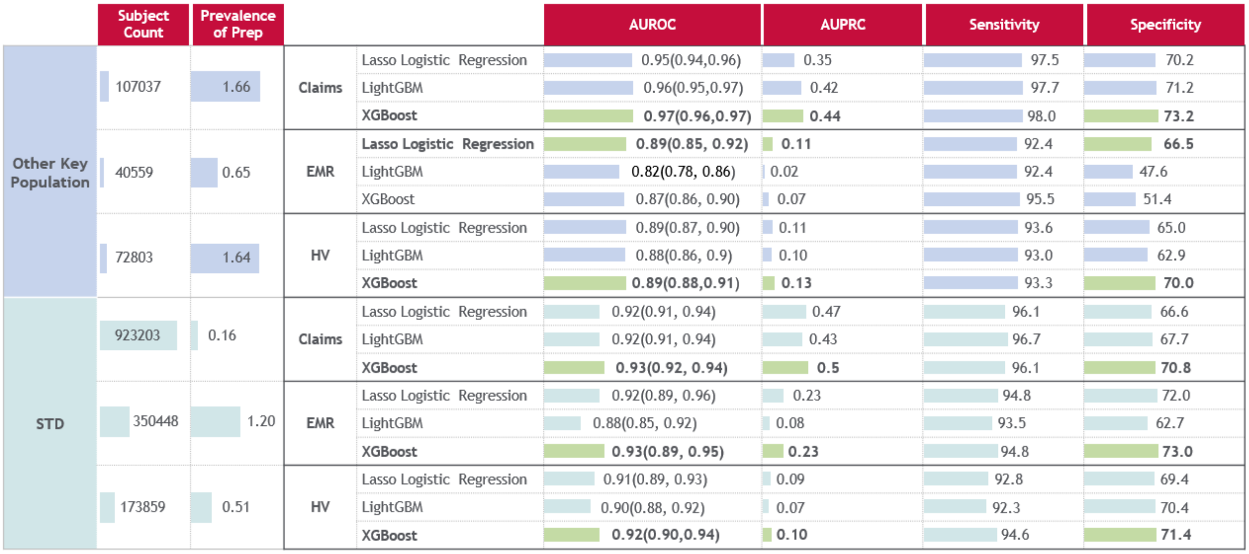
Methods

Individual patient level data were obtained from three data sources in Observational Medical Outcomes Partnership (OMOP) CDM version 5.3: IQVIA PharMetrics Plus Claims (Claims), IQVIA Ambulatory EMR (EMR) and HealthVerity (HV) Marketplace (which includes linked medical records from medical claims, pharmacy claims, EMR, and hospital chargemaster). Between the study period of 2019-01-01 and 2022-06-30, we created two target PWBP cohorts and one outcome cohort of individuals who had been prescribed with PrEP. Individuals with a history of sexually transmitted diseases/infections (STD/STI) were selected in the "STD cohort" and individuals engaging in ICD-10 and CPT codes defined sexual behaviors that were associated with HIV infection, had injectable drug use, or had contacted with and exposed to HIV were included as "Other Key Population". All individuals aged 15 and older on the first occurrence of the above conditions (index date) were included, while those with history of HIV diagnosis/ART treatment, opportunistic infections, chronic hepatitis B, or post-exposure prophylaxis use were excluded.

For predicting the prescription of PrEP among the two PWBP populations, we assessed all individual-level clinical events in the recent past 30 days prior to the index date and during the longer history of up to 2 years prior to the index date. Demographics, comorbidities, past diagnosis, medications, procedures, measurements, and observations were used as covariates for the prediction model. Redundant and infrequent covariates were removed using the FeatureExtraction package from OHDSI Hades and final data were randomly split into 75% training and 25% test sets. Model development was conducted using the Patient Level Prediction (PLP) package from OHDSI Hades with a mix of linear and ensemble boosting algorithms which were trained and evaluated on each data source.[4] Hyperparameter tuning and selection of the best hyperparameters for each model per data source were performed using 3-fold cross-validation. Then the model was evaluated using the model performance on test set and the best model was selected based on the discrimination metrics of area under the receiver operator characteristic curve (AUROC) and area under the precision recall curve (AUPRC). Finally, we analyzed the top predictors of PrEP use as identified by the best predictive model for each data source.

Results

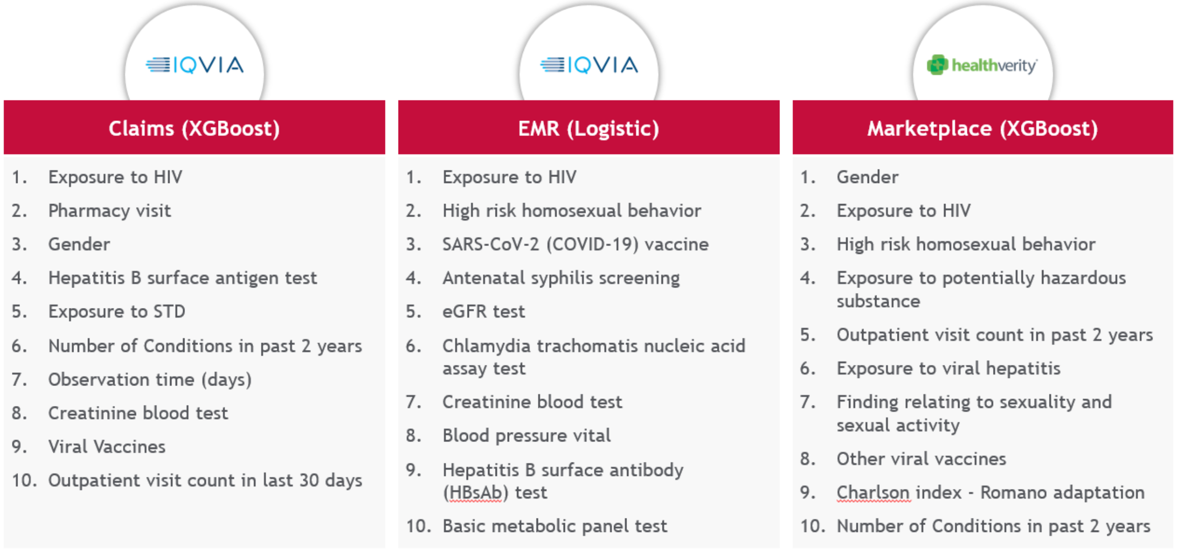
The demographics report of cohort characterization shows the similarities and differences of population distribution in target and outcome cohorts across data sources. Geographic variations in the population were observed, reflecting differences in data sources. For instance, most population in Claims dataset were concentrated in Florida and Texas, while in EMR and HV the majority are from California. The prevalence of PrEP in both cohorts across the three data sources ranged from 0.16% to 1.66%. Lasso logistic regression, XGBoost,[5] and LightGBM[6] models were used for predicting PrEP use and the comprehensive model evaluation results are presented in Figure 1.



	Subject Count	Prevalence of Prep			AUROC	AUPRC	Sensitivity	Specificity
Other Key Population	107037	1.66	**Claims**	Lasso Logistic Regression	0.95(0.94,0.96)	0.35	97.5	70.2
				LightGBM	0.96(0.95,0.97)	0.42	97.7	71.2
				XGBoost	**0.97(0.96,0.97)**	**0.44**	98.0	**73.2**
	40559	0.65	**EMR**	**Lasso Logistic Regression**	**0.89(0.85, 0.92)**	**0.11**	92.4	**66.5**
				LightGBM	0.82(0.78, 0.86)	0.02	92.4	47.6
				XGBoost	0.87(0.86, 0.90)	0.07	95.5	51.4
	72803	1.64	**HV**	Lasso Logistic Regression	0.89(0.87, 0.90)	0.11	93.6	65.0
				LightGBM	0.88(0.86, 0.9)	0.10	93.0	62.9
				XGBoost	**0.89(0.88,0.91)**	**0.13**	93.3	**70.0**
STD	923203	0.16	**Claims**	Lasso Logistic Regression	0.92(0.91, 0.94)	0.47	96.1	66.6
				LightGBM	0.92(0.91, 0.94)	0.43	96.7	67.7
				XGBoost	**0.93(0.92, 0.94)**	**0.5**	96.1	**70.8**
	350448	1.20	**EMR**	Lasso Logistic Regression	0.92(0.89, 0.96)	0.23	94.8	72.0
				LightGBM	0.88(0.85, 0.92)	0.08	93.5	62.7
				XGBoost	**0.93(0.89, 0.95)**	**0.23**	94.8	**73.0**
	173859	0.51	**HV**	Lasso Logistic Regression	0.91(0.89, 0.93)	0.09	92.8	69.4
				LightGBM	0.90(0.88, 0.92)	0.07	92.3	70.4
				XGBoost	**0.92(0.90,0.94)**	**0.10**	94.6	**71.4**

Figure 1. Model results for each target cohort for each data source: IQVIA PharMetrics Plus Claims (Claims), IQVIA Ambulatory EMR (EMR), and HealthVerity Marketplace (HV). Model names in bold are the best performing model for that data source

All the models are fine tuned to have a higher sensitivity at the expense of specificity to ensure the models are accurately predicting PrEP uptake among these individuals. For individuals with other factors associated with HIV infection, XGBoost performed the best in Claims and HV while a simpler lasso logistic regression model outperformed more complex boosting methods in EMR data. In the STD cohort, XGBoost demonstrated the best performance across data sources. Next, the top predictors were analyzed for each of the best models for their respective data sources and the observed top predictors are presented in Figure 2 and Figure 3. For instance, exposure to HIV and gender were two of the top predictors across data sources. However, ICD-10 defined high-risk homosexual behavior was one of the top predictors in EMR and HV but was absent from the covariates used for modeling in Claims data.



Claims (XGBoost)

1. Exposure to HIV
2. Pharmacy visit
3. Gender
4. Hepatitis B surface antigen test
5. Exposure to STD
6. Number of Conditions in past 2 years
7. Observation time (days)
8. Creatinine blood test
9. Viral Vaccines
10. Outpatient visit count in last 30 days

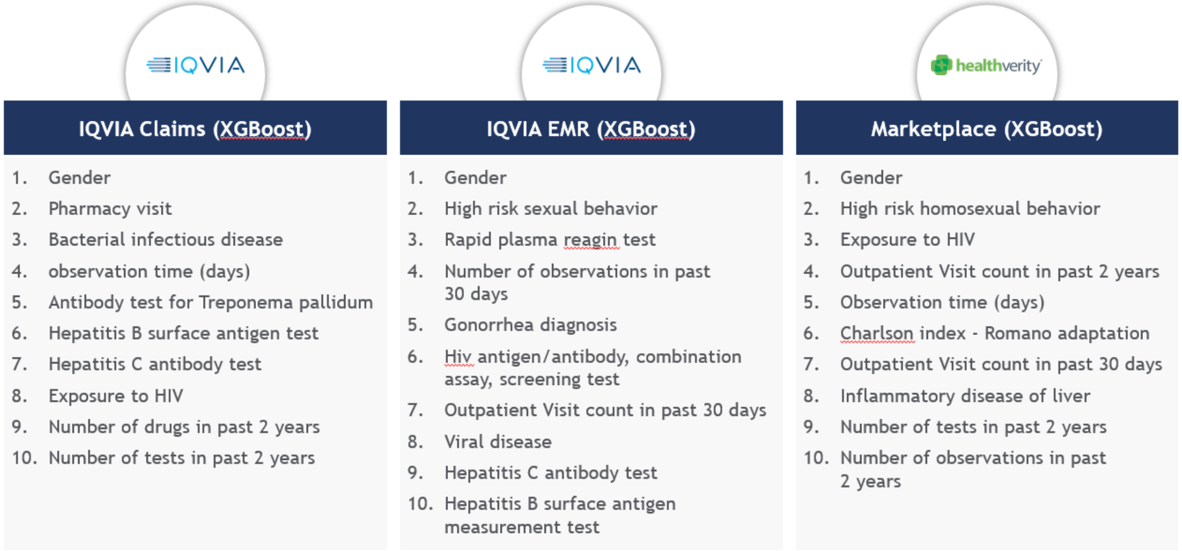
EMR (Logistic)

1. Exposure to HIV
2. High risk homosexual behavior
3. SARS-CoV-2 (COVID-19) vaccine
4. Antenatal syphilis screening
5. eGFR test
6. Chlamydia trachomatis nucleic acid assay test
7. Creatinine blood test
8. Blood pressure vital
9. Hepatitis B surface antibody (HBsAb) test
10. Basic metabolic panel test

Marketplace (XGBoost)

1. Gender
2. Exposure to HIV
3. High risk homosexual behavior
4. Exposure to potentially hazardous substance
5. Outpatient visit count in past 2 years
6. Exposure to viral hepatitis
7. Finding relating to sexuality and sexual activity
8. Other viral vaccines
9. Charlson index - Romano adaptation
10. Number of Conditions in past 2 years

Figure 2. Other Key Population cohort top predictors across data sources



IQVIA Claims (XGBoost)

1. Gender
2. Pharmacy visit
3. Bacterial infectious disease
4. observation time (days)
5. Antibody test for Treponema pallidum
6. Hepatitis B surface antigen test
7. Hepatitis C antibody test
8. Exposure to HIV
9. Number of drugs in past 2 years
10. Number of tests in past 2 years

IQVIA EMR (XGBoost)

1. Gender
2. High risk sexual behavior
3. Rapid plasma reagin test
4. Number of observations in past 30 days
5. Gonorrhea diagnosis
6. Hiv antigen/antibody, combination assay, screening test
7. Outpatient Visit count in past 30 days
8. Viral disease
9. Hepatitis C antibody test
10. Hepatitis B surface antigen measurement test

Marketplace (XGBoost)

1. Gender
2. High risk homosexual behavior
3. Exposure to HIV
4. Outpatient Visit count in past 2 years
5. Observation time (days)
6. Charlson index - Romano adaptation
7. Outpatient Visit count in past 30 days
8. Inflammatory disease of liver
9. Number of tests in past 2 years
10. Number of observations in past 2 years

Figure 3. STD cohort top predictors across data sources

Conclusion

Our study highlights the success of identifying similarities and differences across multiple data sources, facilitated efficiently through the utilization of a LSA platform that incorporates data in OMOP CDM format. Through cohort characterizations, model performance and top predictors for PrEP use, each analysis reinforces the importance of executing studies across diverse data sources to understand the strengths and limitations of the underlying data. Careful examination of study results from multiple data sources is important for generating reliable and generalizable real-world evidence that can inform and guide interventions for effective PrEP utilization in key populations. Additionally, the standardized application and extension of the OMOP CDM ensures the scalability and broad applicability of the findings, further shaping PrEP implementation strategies to diverse populations defined by demographics and geographies globally. As next steps, we want to evaluate this method's effectiveness to facilitate feasibility analysis in selecting the right data source for the research question of the real-world study.

References

1. Taur SR. Observational designs for real-world evidence studies. Perspect Clin Res. 2022 Jan-Mar;13(1):12-16. doi: 10.4103/picr.picr_217_21. Epub 2022 Jan 6. PMID: 35198423; PMCID: PMC8815667.
2. Mayer KH, Agwu A, Malebranche D. Barriers to the Wider Use of Pre-exposure Prophylaxis in the United States: A Narrative Review. Adv Ther. 2020 May;37(5):1778-1811. doi: 10.1007/s12325-020-01295-0. Epub 2020 Mar 30. PMID: 32232664; PMCID: PMC7467490.
3. Centers for disease Control and Prevention. PrEP for HIV Prevention in the U.S. Newsroom. US [updated November 23, 2021] Available from: https://www.cdc.gov/nchhstp/newsroom/fact-sheets/hiv/PrEP-for-hiv-prevention-in-the-US-factsheet.html
4. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek P (2018). “Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data.” Journal of the American Medical Informatics Association, 25(8), 969-975. https://doi.org/10.1093/jamia/ocy032.
5. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785
6. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., … Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146–3154.