Predictive Modeling for Early Identification of Pulmonary Arterial Hypertension Using Real-World EHR and Claims Data

Saurabh Pandey, **Pankaj Bhardwaj,** Vikash Verma, Louis Brooks Jr, Marissa Seligman, Shashi Khan, Abhimanyu Roy, Abhinav Nayyar, Ankit Arora, Ankita Misra, Rajeev Kumar, Varshith Gandla, Riddhi Markan, Vaibhav Bansal, Anuj Gupta, Vishan Khatavkar, Kavita Karayat, Aakash Singh, Srishti Motila, Gargi Mahashay

Background

- Pulmonary Arterial Hypertension (PAH) is a rare, progressive disease affecting 2.28 per 100,000 patients[1,2]
- Early diagnosis remains challenging as the patient symptoms are non-specific often similar to cardiorespiratory disorders, thereby resulting in prolonged diagnostic delays of up to 4 years and reliance on invasive testing such as right heart catheterization (RHC) for disease confirmation.[3,4]
- Previous claims-based machine-learning algorithm have showed feasibility for identifying PAH patients, 6-months prior to diagnosis, using U.S. administrative claims data.[1]
- Earlier identification may enable timely treatment initiation and reducing downstream healthcare resource utilization and costs, and disease progression.

Objective

- Develop and evaluate a machine learning model predicting PAH using routinely captured data from real world claims and EHR data to identify patients at elevated risk of developing PAH prior to invasive confirmatory testing.

Methodology

Study Design & Data Source:
A retrospective case–control study was conducted using the Optum® Market Clarity database spanning 01 July 2020 through 30 June 2025.

Case Identification:
Pulmonary Arterial Hypertension (PAH) cases were identified based on ≥1 medical claim with a PAH diagnosis using ICD-10-CM diagnosis code I27.0. The first qualifying PAH claim was defined as the index date.

Clinical Confirmation & Eligibility:

- Patients were required to be ≥18 years of age
- No prior history of pulmonary arterial hypertension (PAH)
- Continuous health plan enrollment in claims data with documented clinical activity in the EHR during the 48-month baseline period prior to the index date
- Evidence of right heart catheterization (RHC) during the 12-month lookback period before the index date

Matching Strategy:
Non-PAH controls were selected from individuals without any PAH diagnosis and were matched at a 1:2 ratio to PAH cases based on age, gender, race, geographic region, and Charlson Comorbidity Index (CCI) using propensity score matching.

Feature Extraction:
Baseline variables derived from both claims and EHR data were captured during the 48-month baseline period, including

- Comorbid conditions: Cardiac structural abnormalities, valve disorders, disorders of the circulatory system, depression and anxiety.
- Clinical Manifestations: Abnormal breath sounds, dyspnea, chest pain, hypoxia, generalized weakness, rhonchi, respiratory failure, pulmonary congestion
- Diagnostic testing patterns, Healthcare utilization measures.

Model Development:
An 80/20 train–test split was used to develop and evaluate predictive models. Logistic Regression (LR), Random Forest (RF), and XGBoost algorithms were trained using baseline features to predict PAH status. Model performance was assessed using the area under the receiver operating characteristic curve (AUC) and F1 score.

Results

- A total of 1,026 PAH patients (342 PAH cases and non-PAH controls (N=684) were included, with over 60% aged ≥70 years and ~80% were Caucasian.
- Recursive feature elimination was utilized to iteratively removed predictors with near-zero logistic regression coefficients. The top 15 most informative features for PAH prediction were retained for the modelling.
- All three models showed comparable predictive performance –with F1 scores of 77% (LR), 79% (XGBoost), and 79% (RF). (Figure 1)
- Dyspnea, ECG abnormalities, non-ST elevation myocardial infarction (NSTEMI), pulmonary congestion, and cardiac abnormalities were the top claims-EHR predictors for PAH diagnosis (Figure 2).

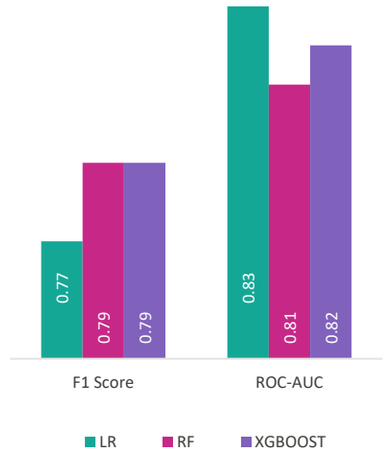

Figure 1. Model Comparison

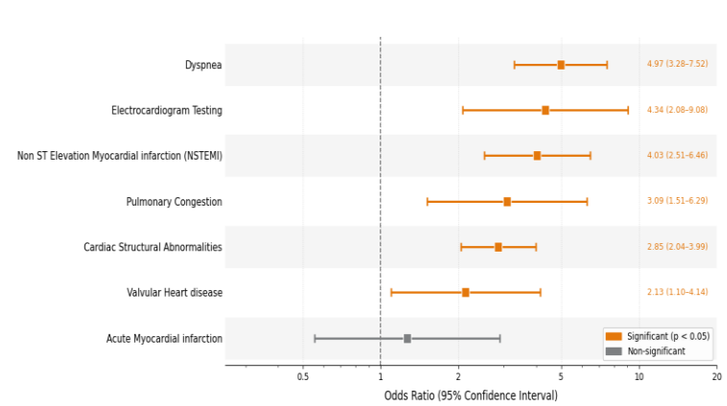

Figure 2: Predictors of PAH using Logistic Regression Odds Ratio (95% CI)

Conclusions

- This predictive model demonstrated the ability to identify patients at elevated PAH risk up-to-48 months prior to diagnosis, building on prior literature. These findings support the feasibility of leveraging integrated EHR and claims data for earlier identification of at-risk PAH patients.
- A key limitation was aggregation of features across the 48-month baseline, limiting differentiation of early versus late risk signals.
- Future analyses using time-windows in the baseline period to help distinguish early and late predictors.

References 1. Hyde B, et al. Pulm Circ. 2023;13(2):e12237; 2. GBD 2021 Collaborators. Lancet Respir Med. 2025. PAH burden, 1990–2021; 3. Didden EM, et al. Pulm Circ. 2023:13(1);e12188; 4. Rosenberg J, et al. Am J Manag Care. 2023;

Optum