Evaluation of Healthcare Insurance Claims Record based Artificial Intelligence Screening Tools for Undiagnosed Obstructive Sleep

Sam Rusk, BS, (1) Fred Turkington, BS, (1) Chris R. Fernandez, MS, (1) Yoav N. Nygate, MSc, (1) Nick Glattard, MS, (1) Melania Abrahamian, PhD, (1) Tom Vanasse, PhD, (1) Dana Richardson, BSN, MAB, (2) Tim Bartholow, MD, (2) Nathaniel Watson, MD, MSc (3)

Introduction

Obstructive sleep apnea (OSA) identified by recurrent partial and complete upper airway obstructive events affects about 34% and 17% of middle-aged men and women respectively1. Researchers estimate that about 80%-90% of people in US with OSA are undiagnosed and OSA is estimated to affect >50 million adults in the U.S. and North America.

The prevalence of sleep apnea is increasing in the general population in association with risk factors including obesity, aging, and comorbidity. OSA exacerbates costly co-morbid conditions and leads to increased adverse health risks and those associated costs over time. Despite the rise in importance of sleep wellness as contributor to human health, the accessibility of accurate OSA diagnostic tools, OSA treatment options and their cost are still considered as major challenges. (1)

Therefore, we are introducing the application of our AI screening approach that uses insurance claims as input to contribute to the efforts improving the screening methods hence helping with the reduction of costs through early and proper identification of individuals with undiagnosed OSA. Accordingly, we demonstrate the AI impact in personalized healthcare.

Methodology

Input data was extracted from a dataset composed of medical and pharmacy claims (years 2016 – 2020) from the Wisconsin All-Payor claims database including coverage of >4,000,000 patients, >10,000 ICD codes, and >$50 billion in medical spending. A total of 2,431,852 patients and 39,712 unique federal drug identification codes were included within 91.5 million pharmacy claims in the dataset. In addition, we identified 94 drug groups from the above-mentioned identified drugs. Moreover, demographics information (age and gender) was extracted for the patients to account for their contribution to the patient group (OSA positive, OSA negative) classification.

The retrieved data were used to construct input features by counting the total number of claims for each unique drug or drug groups in each subject resulting in a patient-level feature vector of 39,712 drug frequencies per medication or 94 drug groups. The positive sleep apnea population was defined by individuals who had both at least one medical claim for sleep apnea diagnosis (ICD codes G4733/G4731) and an appropriate sleep test (CPT codes 9580*/9581*/99213).

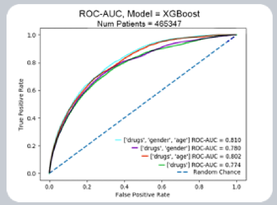
Methodology Continued

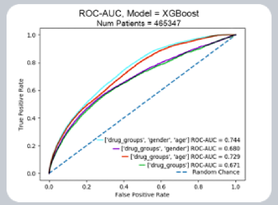
Two models, Logistic Regression and XGBoost were used for detecting OSA. Each model performance was evaluated by ROC-AUC statistics through cross validation (CV) approach, where 10-fold CV was implemented using 2,094,059 or 2,094,058 and 2,326,73 or 2,326,74 patients for train and validation datasets, respectively. The final model performance was reported on 1,861,385 (80% population) and 4,653,47 (20% population) patients as train and test datasets, respectively. We performed model explainability analysis through SHAP and applied both global and local SHAP analyses, which resulted in the identification of top features that contribute to the classification of patients as OSA positive group.

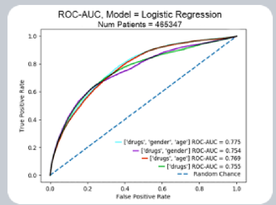
Results

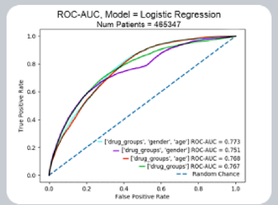
Model Performance Results: Logistic Regression and XGBoost models detecting OSA based on observed medication frequencies and demographics features indicate that models that use both demographics features (age and gender) along with the individual medications have the best performance.

In addition, our results show that XGBoost performs better compared to Logistic Regression when individual medications are considered. Moreover, our findings indicate that performance of XGBoost is improved as one or both demographics features are included with drug groups as input features, while Logistic Regression shows the performance improvement once both demographics features, and drug groups are considered.









Results Continued

Model Explainability Results: SHAP Linear Explainer and TreeExplainer were used for model explainability of Logistic Regression and XGBoost, respectively.

The analyses using individual medications and drug groups indicate that at global level diabetes, hypertension, and hyperlipidemia medications are among the top features that pull the classifiers towards OSA positive classification.

And the local SHAP analysis support the global SHAP results indicating that patients using combination of medications such as antidiabetics and antihypertensives or combination of antidiabetics, antihypertensives and antihyperlipidemics are classified as patients with OSA.

Individual Medications

Drug Name	Drug Groups	SHAP values	Drug Name	Drug Groups	SHAP values
Estradiol	Estrogens	0.51	Levothyroxine Sodium	Thyroid Agents	-0.251
Terazosin Hcl	Antihyper-tensives	0.494	Duloxetine Hcl	Antidepressants	-0.269
Triamterene & Hydrochlorothiazide	Antihyper-tensives	0.456	Hydrocodone-acetaminophen	Analgesics - Opioid	-0.376
Eplerenone	Antihyper-tensives	0.341	Pancrelipase (Lipase-protease-amylase)	Digestive Aids	-0.410
Sildenafil Citrate	Cardiovascular Agents - Misc.	0.279	Fluticasone Propionate (Nasal)	Nasal Agents Systemic And Topical	-0.468
Simvastatin	Antihyper-lipidemics	0.100	Albuterol Sulfate	Antiasthmatic and Bronchodilator Agents	-1.048

Drug Groups

Drug Groups	SHAP values	Drug Groups	SHAP values
Allergenic Extracts/Biologicals Misc	0.485	Dermatologicals	-0.133
Neuromuscular Agents	0.251	Penicillins	-0.133
ADHD/Anti-narcolepsy/Anti-obesity/Anorexiants	0.189	Musculoskeletal Therapy Agents	-0.154
Antiparkinson And Related Therapy Agents	0.154	Anticonvulsants	-0.237
Antihyperlipidemics	0.100	Analgesics - Anti-inflammatory	-0.259
Antidiabetics	0.900	LOCAL Anesthetics-parenteral	-0.417

Results Continued

The tables summarize the top features identified by SHAP values, where positive values are the top features for patients classified as OSA positive while negative values are the top features that pull model towards OSA negative group classification.

In addition, SHAP local analysis highlighted the contribution of age to OSA positive classification along with the identified medications.

Conclusion

In this research study we demonstrate health insurance claims records contain predictive information that can aid in more systematic screening of undiagnosed conditions like OSA.

Furthermore, in a statistical analysis of feature importance, we observed medications indicative of comorbidities with known association to OSA.

In addition, we identified that demographic features (age and gender) are also among the top features and with the combination of the identified medication contribute to our improved AI screening approach.

These findings are useful to clinicians and payers in identifying undiagnosed OSA populations, including those responsible for value-based payment models.

Future Work

The presented AI research approach has the capability for screening other undiagnosed diseases such as neurodegenerative diseases using the health insurance claims records.

We expect our AI based approach support clinicians and patients regarding other diseases diagnostic predictions to contribute to ongoing efforts in personalized healthcare.

Authors

1. EnsoData Research, EnsoData, Madison, WI, USA

2. Wisconsin Health Information Organization

3. Department of Neurology, University of Washington School of Medicine, Seattle, WA, USA