PCR265

Incorporating Social Determinants of Health (SDOH) Into a Screening Tool Using Machine Learning to Predict Lung Cancer Diagnosis

Bianca Jackson, MPH[1,2], Charisse Madlock-Brown, PhD[3]

[1]OPEN Health HEOR & Market Access, New York, NY, USA Contact: biancajackson@openhealthgroup.com.
[2]University of Tennessee Health Science Center, Memphis, TN, USA. [3]University of Iowa, Iowa City, IA, USA

OPEN HEALTH

INTRODUCTION

- Lung cancer is one of the most common diseases in both men and women and is the leading cause of cancer-related deaths.[1]
- The timing of a lung cancer diagnosis is essential to patient prognosis and survival, considering that late diagnosis of lung cancer contributes to increased mortality risks.[2]
- Recent studies and investigations "suggest that using lung cancer risk prediction models could lead to more effective screening programs compared to the current recommendations."[3]
- However, current models do not take into account social risk factors.

OBJECTIVES

- This study aimed to develop a predictive screening tool, incorporating social determinants of health (SDOH) data, to improve model accuracy in the earlier identification of patients at risk of developing lung cancer. The study had the following aims:
  - Aim 1: Examine the association between social determinants and lung cancer incidence.
  - Aim 2: Establish a screening tool for high-risk patients.
  - Aim 3: Assess models for algorithmic fairness.

METHODS

Study Design and Data Source

- This study employs an observational, retrospective cohort design and uses de-identified EHR data from the Research Enterprise Data Warehouse (rEDW), which covers patients from three institutions in the state of Tennessee.
- Baseline data from 2018 and 2019 are used to predict lung cancer diagnosis in 2020.
- Lung cancer diagnosis is based on ICD-10 codes with thoracic or lung neoplasms.

Patient Population

- Study participants must be considered "high risk", defined as:
  - at least 50 years of age
  - having a smoking history, and
  - no prior history of lung cancer.

Predictor Variables

- Extracted patient-level characteristics include demographic information and clinical parameters, such as comorbidities, symptoms, procedures, and laboratory test results (**Table 1**).
- Additional features are identified from relevant literature.
- SDOH features within four SDOH domains (socioeconomic status (SES), household characteristics, racial & ethnic minority status, and housing type & transportation) are captured from the American Community Survey (ACS) (2018-2019).
- Patient home addresses are geocoded and linked with the ACS data to obtain SDOH data at the census-tract and county level.

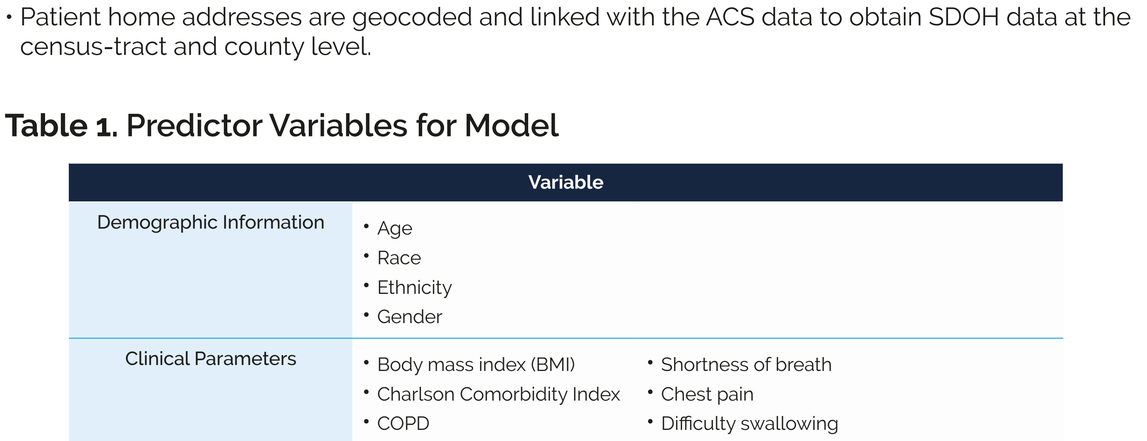
Table 1. Predictor Variables for Model

Variable		
Demographic Information	• Age • Race • Ethnicity • Gender	
Clinical Parameters	• Body mass index (BMI) • Charlson Comorbidity Index • COPD • Emphysema • Hay Fever • Asthma • Anxiety • Fatigue • Cough	• Shortness of breath • Chest pain • Difficulty swallowing • Hypertension • Coronary artery disease • Angina pectoris • Stroke • Diabetes • Chronic bronchitis • Kidney failure

RESULTS

- The study population includes a total of 46,470 patients, of whom approximately 1,400 patients had a lung cancer diagnosis.

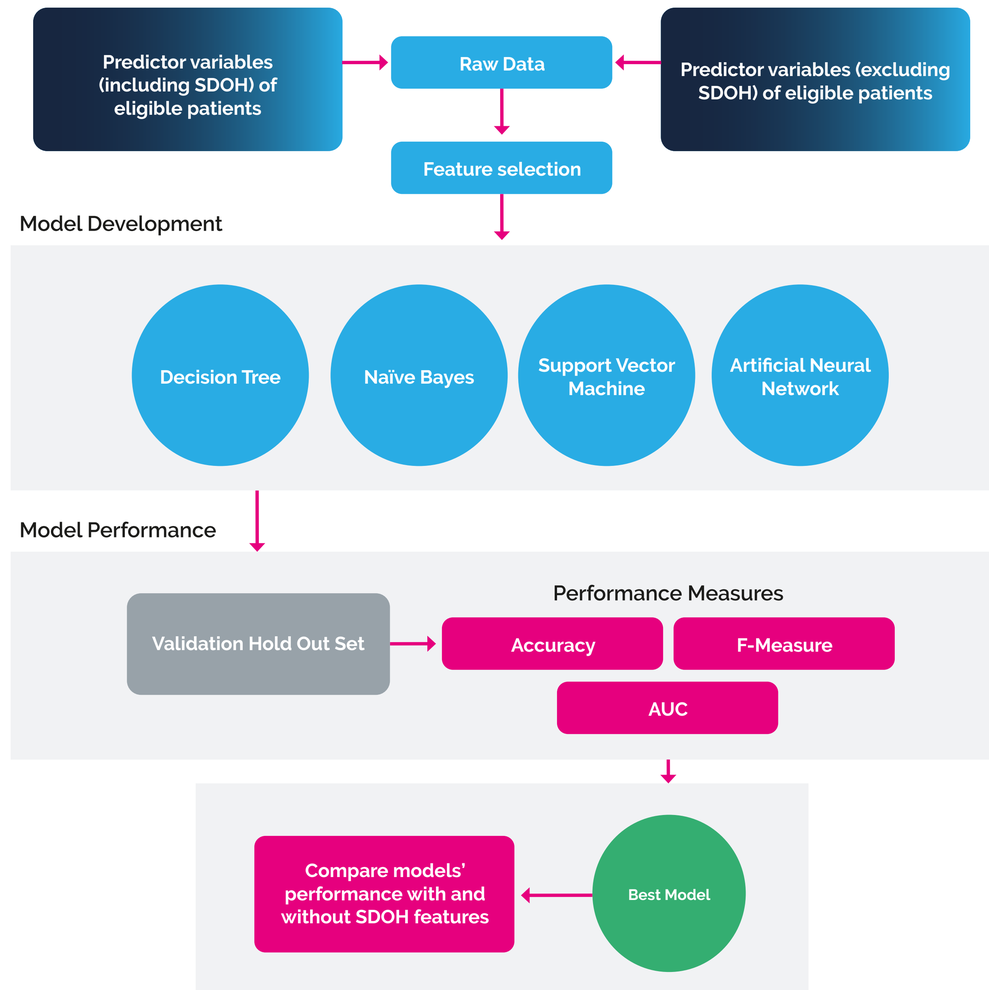
Aim 1. Examine the association between social determinants of health and lung cancer incidence



The stepwise regression model is compared with the baseline model, by adding SDOH domains, to understand where considerable gain is observed.

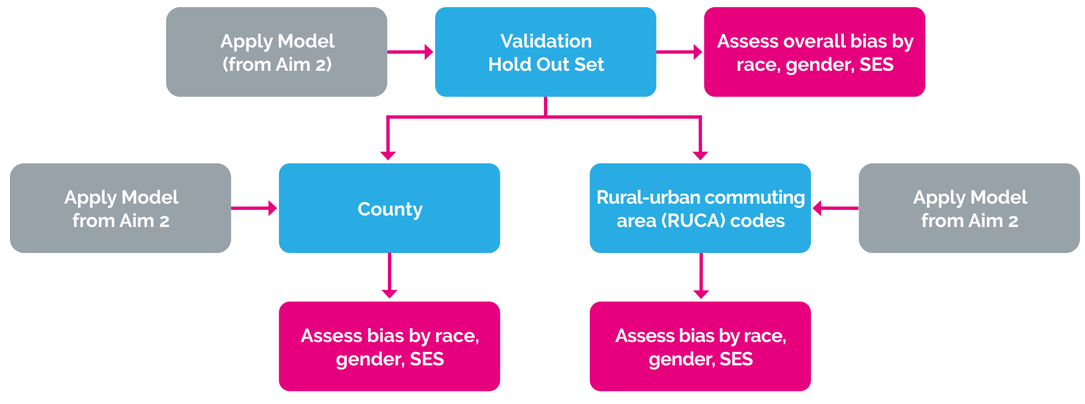
RESULTS (CONT)

Aim 2. Establish a screening tool for high-risk patients



1. Feature Selection: A forward selection algorithm, which captures features, through speculative rounds, that increase the performance of the model is used.
2. Model development: These features are applied to decision tree, naïve Bayes, support vector machine, and artificial neural networks models.
3. Model performance: The raw data are split using a validation holdout set for testing to assess each model's performance on unseen data. Accuracy, AUC, and F-measure metrics are used to examine model performance. The best model is determined by assessing performance measures where sensitivity is maximized without a significant drop in specificity.

Aim 3. Assess model for algorithmic fairness



1. The model containing SDOH developed from Specific Aim 2 is applied on a validation holdout set to assess overall bias by race, gender, SES.
2. The validation set is further stratified by county and rural-urban commuting area (RUCA) codes. RUCA codes classify U.S. census tracts using measures of population density, urbanization, and daily commuting. The model is applied to these subsets of data and further assessed for bias by race, gender, and SES.

CONCLUSIONS

- The work outlines how a predictive screening tool incorporates more inclusive criteria for predicting lung cancer while using data that is routinely collected in EHR databases and can be feasibly incorporated into clinical settings.
- Early identification of patients at high risk for developing lung cancer allows for clinicians to efficiently implement recognized screening guidelines, identify lung cancer at earlier stages, and ultimately, decrease adverse health outcomes amongst patients with the highest risk of developing lung cancer.
- This research study highlights how patient SDOH differences contribute to disparities in lung cancer risks. Incorporating SDOH into screening tools is one step closer to achieving NCI's National Cancer Plan's goal to eliminate inequities.

REFERENCES

1. American Cancer Society. Key Statistics for Lung Cancer. 2024. Retrieved from: https://www.cancer.org/cancer/types/lung-cancer/about/key-statistics.html; **2.** Gray, E. P., Teare, M. D., Stevens, J., & Archer, R. (2016). Risk Prediction Models for Lung Cancer: A Systematic Review. Clin Lung Cancer, 17(2), 95-106; **3.** Haaf KT, Jeon J, Tammemägi MC, Han SS, Kong CY, Plevritis SK, Feuer EJ, de Koning HJ, Steyerberg EW, Meza R. Risk prediction models for selection of lung cancer screening candidates: A retrospective validation study. (2017). PLOS Medicine.14(4):1-24.

ACKNOWLEDGMENTS

We are grateful to The University of Tennessee Health Science Center for the provision of data from the Research Enterprise Data Warehouse (rEDW) dataset.

Presented at: ISPOR Europe 2024