TCORS 2.0
University of Michigan & Georgetown University

Center for the Assessment of Tobacco Regulations [CAsToR]

Are the relevant risk factors being adequately captured in empirical studies of smoking initiation? A machine learning analysis based on the PATH study

Thuy T. T. Le[1*] PhD, Mona Issabakhsh[3] PhD, Yameng Li[3] MS, Luz María Sánchez-Romero[3] PhD, Jiale Tan[2] MS, Rafael Meza[4] PhD, David Levy[3] PhD, David Mendez[1] PhD

Introduction

Cigarette smoking continues to pose a threat to public health. Identifying individual risk factors for smoking initiation is essential to further mitigate this epidemic. To our knowledge, no study today has used Machine Learning (ML) techniques to automatically uncover informative predictors of smoking onset among adults using the Population Assessment of Tobacco and Health (PATH) study.

Methods

In this work, we employed Random Forest paired with Recursive Feature Elimination to identify relevant PATH variables that predict smoking initiation among adults who have never smoked at baseline between two consecutive PATH waves. We included all potentially informative baseline variables in wave 1 (wave 4) to predict past 30-day smoking status in wave 2 (wave 5). Using the first and most recent pairs of PATH waves was found sufficient to identify the key risk factors of smoking initiation and test their robustness over time. The eXtreme Gradient Boosting method (XGBoost) was employed to test the quality of these selected variables.

Results

As a result, classification models suggested about 60 informative PATH variables among many candidate variables in each baseline wave. With these selected predictors, the resulting models have a high discriminatory power with the area under the Specificity-Sensitivity curves of around 80% (see Figure 1). We examined the chosen variables and discovered important features. Across the considered waves, two factors, (i) BMI and (ii) dental/oral health status, robustly appeared as important predictors of smoking initiation, besides other well-established predictors.

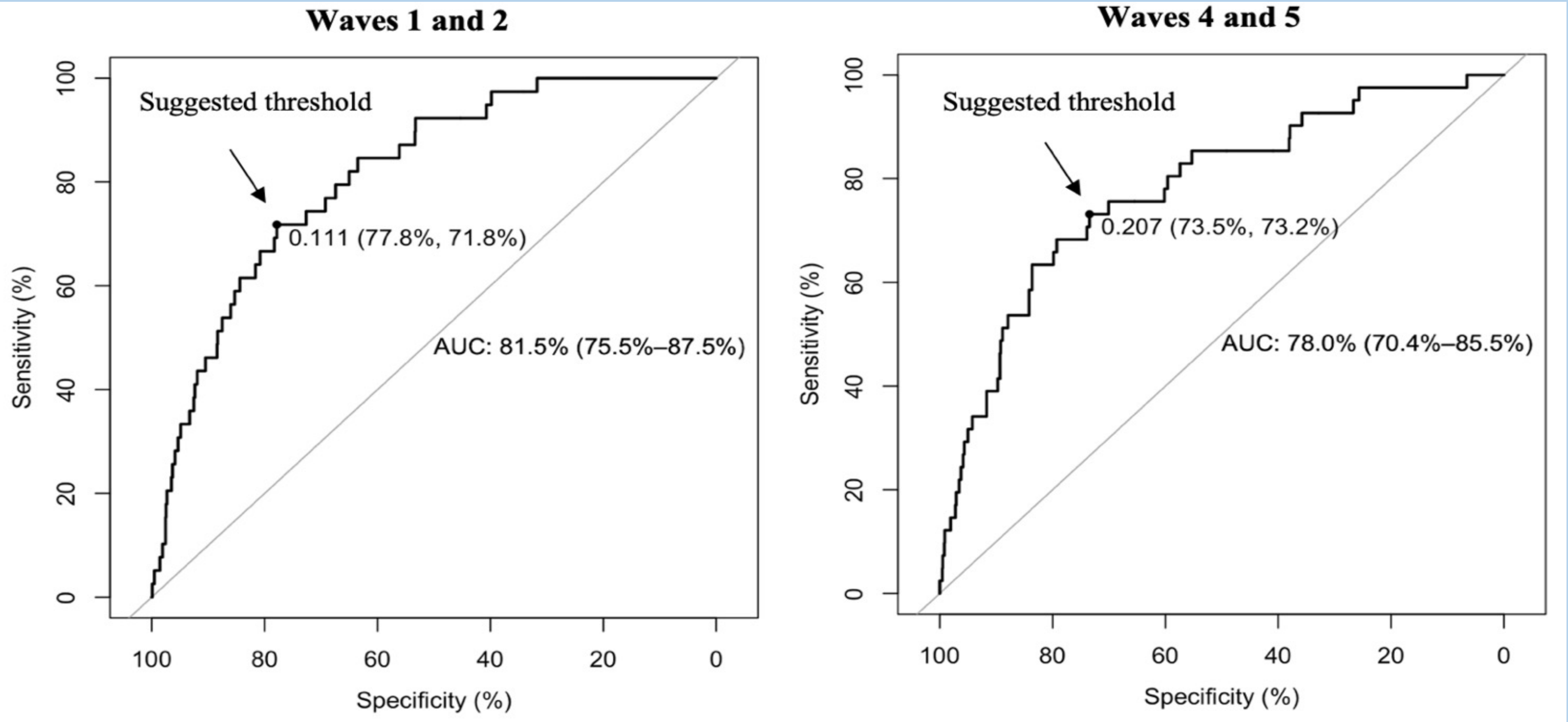

Figure 1: The ROC curves of all the XGBoost classifiers. The diagonal line in each plot represents random guessing. Each plot shows the AUC curve (95% CI) together with the optimal threshold (Specificity, Sensitivity).

Discussion

- Our work demonstrates that ML methods are useful to predict smoking initiation with high accuracy, identify novel smoking initiation predictors, and to enhance our understanding of tobacco use behaviors.
- Understanding individual risk factors for smoking initiation is essential to prevent smoking initiation. With this methodology, a set of the most informative predictors of smoking onset in the PATH data was identified. Besides reconfirming well-known risk factors, the findings suggested additional predictors of smoking initiation that have been overlooked in previous work. More studies that focus on the newly discovered factors (BMI and dental/oral health status,) are needed to confirm their predictive power against the onset of smoking as well as determine the underlying mechanisms.

Author Affiliations:

- [1]University of Michigan, School of Public Health, Department of Health Management and Policy, Ann Arbor, MI, USA. *Corresponding email: thuyttle@umich.edu
- [2]University of Michigan School of Public Health, Department of Epidemiology, Ann Arbor, MI, USA
- [3]Georgetown University-Lombardi Comprehensive Cancer Center, Washington, DC, USA
- [4]Integrative Oncology, BC Cancer Research Institute, Vancouver BC

UNIVERSITY OF MICHIGAN
GEORGETOWN UNIVERSITY
Yale
SCHOOL OF PUBLIC HEALTH
UNIVERSITY OF MICHIGAN

Support is provided by grant U54CA229974 from the National Institutes of Health, National Cancer Institute and Food and Drug Administration (FDA)