Are the relevant risk factors being adequately captured in empirical studies of smoking initiation? A machine learning analysis based on the Population Assessment of Tobacco and Health study

Thuy Le et al.

Research question

- Using the PATH survey, the research question is to predict smoking initiation between two consecutive waves among adult never smokers at baseline and identify relevant predictors of this behavior.

- Baseline population: individuals who never tried smoking even one or two puffs of a cigarette - never smokers.
- Target population: individuals did/did not smoke a cigarette in the past 30 days (the past 30-day (P30D) smoking status).

Data

- Analyzed the earliest (waves 1 and 2) and more recent (waves 4 and 5) pairs of waves from the PATH data
- Removed non-relevant variables (e.g., personal identity numbers, random questions, sample weights, and imputed variables).
- Furthermore, variables with more than 5% missing values of the total sample size were dropped to maintain the highest possible number of predictors.
- Excluded individuals with missing smoking status in the outcome waves.

Data

Baseline smoking status	Outcome smoking status	Value of outcome smoking status		Number of predictors	Total
		Yes	No		
Never smokers in wave 1	P30D smoking status in wave 2	197 (3.4%)	5579 (96.6%)	209	5776 (100%)
Never smokers in wave 4	P30D smoking status in wave 5	208 (2.6%)	7687 (97.4%)	293	7895 (100%)

Table 1: A description of the clean and complete datasets extracted from the PATH data after data processing.

Statistical analysis

- For each pair of waves, we used an RF classifier combined with RFE to obtain a subset of predictors on which the RF classifier performs best
- Trained a RF using the selected predictors

Results

- RF- RFE selected a list of only about 60 relevant variables

- Across the considered waves, three factors, (i) BMI, (ii) dental/oral health status, and (iii) taking anti-inflammatory or pain medication, robustly appeared as significant predictors of smoking initiation, besides other well-established predictors.

Results

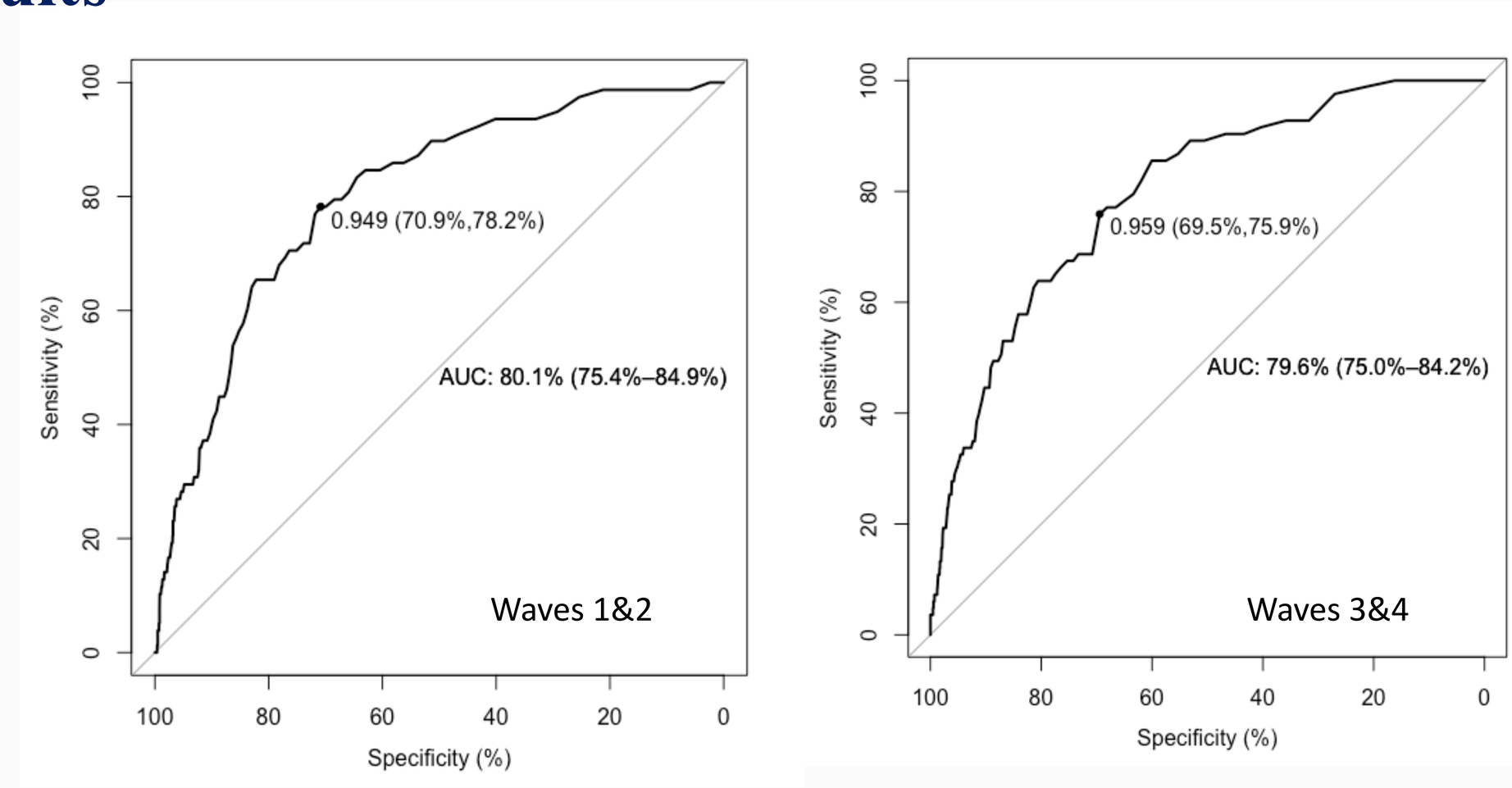

Figure 1: The ROC curves of all the RF classifiers.

Discussion

- Leverage all possible original public adult PATH variables to predict the transition from never to P30D cigarette smokers between two consecutive waves using RF-RFE

- The model performs well in classifying smoking status among never smokers (AUC ≈ 80%)

- About 60 variables associated with the smoking onset between two considered PATH waves among adult

- BMI, dental/oral health status, and taking anti-inflammatory or pain medication, have robustly appeared as significant predictors of smoking initiation

Impact of pilot project

- Two submitted articles
- Having a chance to pursue this research direction
- Producing some preliminary results for future funding applications

Acknowledgments

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