Using statistical machine learning to find complex interactions and important CVD risk factors when predicting general health in adults

Peter D Hart, PhD

Background: Cardiovascular disease (CVD) is the leading cause of premature mortality among U.S. adults. Many risk factors for CVD are established and widely used in health promotion and preventive medicine. However, the extent to which the major CVD risk factors interrelate in relation to health outcomes is less understood. The purpose of this study was to use statistical machine learning to identify complex interactions and important variables when predicting general health (GH) with CVD risk factors. **Methods**: The analysis plan included five objectives. First, a decision (regression) tree was built on training data and fine-tuned using validation data. Second, ordinary least squares (OLS) regression was used to confirm terminal splits provided by the decision tree algorithm. Third, new test data were used to evaluate generalization of the decision tree branches. Fourth, a random forest was run and examined for consistency with decision tree fit performance using training, validation, and test data. Fifth, CVD risk factor variable importance was assessed along with a sensitivity analysis to examine stability in rankings. The 2017-2018 NHANES (N = 3,487) was used for training and validation and 2015-2016 NHANES (N = 3,897) for testing. A residualized self-assessed GH T-score with age, race/ethnicity, sex, and income removed, served as the outcome variable (aka., target). Eight CVD risk factors inspired by Life's Essential 8 (LE8) were used as predictors (aka., features or inputs) and included healthy eating index (HEI; 0-100), moderate-to-vigorous-physical activity (MVPA; min/week), nicotine exposure (NE; non-smoker, quit smoker, other nicotine device user, smoker), sleep time (ST; hr/day), body mass index (BMI; kg/m^2), non-high density lipoprotein cholesterol (NHDL; mg/dL), glycohemoglobin (A1C; %), and mean arterial pressure (MAP; mmHg). SAS HPSPLIT and HPFOREST were the primary reporting procedures. The variable importance sensitivity analysis was performed using R (train and randomForest) and Python (DecisionTreeRegressor and RandomForestRegressor). **Results**: The decision tree built on training data and 10-fold cross validation resulted in a 16-leaf tree with a 6-node depth ($ASE_{Training}$ = 86.3, $ASE_{Validation}$ = 91.1, Δ = 5.6%). BMI split first with A1C splitting next for high BMI (BMI $\geq$ 30.1) and MVPA splitting next for low BMI (BMI < 30.1). OLS regression confirmed (ps < .05) all terminal splits in the training data. Greatest GH (Mean = 56.2) was observed in those with low BMI (BMI < 30.1), high MVPA (MVPA $\geq$ 137.2 min/day), low NE (non-smoker, quit smoker, other nicotine device user), low A1C (A1C < 6.2%), and high HEI (HEI $\geq$ 37.8). Lowest GH (Mean = 43.8) was observed in those with high BMI (BMI $\geq$ 30.1) and high A1C (A1C $\geq$ 6.8%). The decision tree generalized well (ASE_{Test} = 93.1, Δ = 8.0%) with OLS regression confirming (ps < .05) majority of terminal splits. Decision tree variable importance rankings were consistent with the random forest ($r_{Spearman}$ = .83, p = .011) and robust against the sensitivity analysis (avg $r_{Spearman}$ = .84, p = .009, ICC(3,6) = 0.97). **Conclusion**: This study demonstrated a novel use of machine learning that complements conventional statistical analyses. Decision trees along with random forests can identify extremely complex patterns in data and identify variables that contribute the most to group separation of an outcome variable. BMI, MVPA, A1C, and NE are likely the more important predictors of GH in this population.

Keywords: Data science, Machine learning, Cardiovascular disease (CVD), Population health