Men are predictable: Modeling cardiovascular disease prevalence from population survey data

Jason A. Davis, Rhodri Saunders

1. Coreva Scientific, Freiburg, Germany.

OBJECTIVES: Calculating the economic burden of disease requires data regarding disease prevalence. National estimates can be derived from surveys of the general population, which may also access individuals not actively participating in the healthcare system. The Behavioral Risk Factor Surveillance System (BRFSS) is the largest annual country-wide population sampling of health and risk factors. The fidelity of these data, however, may be questionable, relying on accurate self reporting. Cardiovascular disease (CVD) prevalence was examined by gender to assess the feasibility of predicting future trends. **METHODS:** BRFSS data were trimmed to complete cases for 9 CVD risk factors: gender, age, race, overweight, physical activity, diabetes, high blood pressure, smoking and alcohol consumption. Data from 2011 and 2013 were used to train Bayesian and tree-based algorithms to evaluate predictor performance on unseen data from subsequent years (2013 and 2015) by comparing predicted with reported prevalence. **RESULTS:** For algorithms used, predictions of future prevalence were significantly better for males than females ($p < 0.001$, Šidák multiple testing correction). In the best performing algorithm (Naïve Bayes), the mean percent difference from the actual prevalence for males was 3.8±2.5% and females 151±62% ($p < 0.05$, two-tailed t-test). Data from 2013 yielded better 2-year predictions (2015) for women than the same time span with 2011 data (2011 to 2013, $p < 0.05$, two-tailed t-test), while for men, there was no significant difference ($p = 0.54$, two-tailed t-test). Models trained on the genders combined resulted in underestimates of prevalence ($p < 0.001$, Z-test). **CONCLUSIONS:** Patient-reported survey data can be used to predict cardiovascular disease prevalence. Accuracy of estimation is better in males versus females. Given that BRFSS data are retrospective, our findings may reflect more substantial lifestyle changes in females or suggest discussion on changes in how survey data from female respondents are collected.

Introduction

- Accurate future projections of disease prevalence rely on accurate sampling of current trends in the general population
- Other models of long-term prediction of cardiovascular disease [CVD][1,2] have focused on estimates of mortality and costs
- Medical records have greater detail, but will only sample individuals presented to the health system
- General survey data can reach a broader sampling of the population, but may be less detailed
- An analysis was performed to assess feasibility of incorporation of more risk factors from retrospective self-reported data to model future CVD

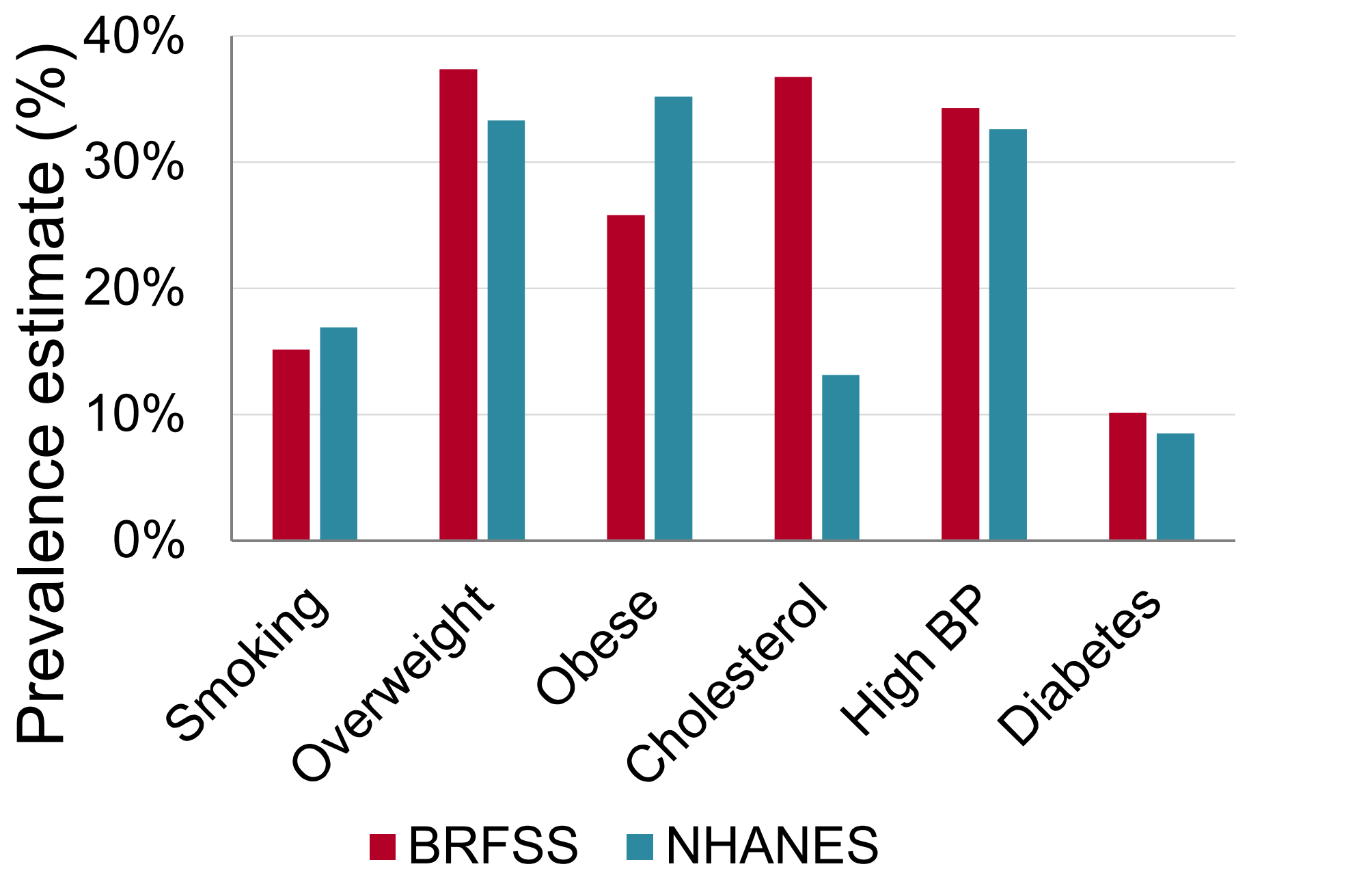
Methods

- Analysis of Behavioral Risk Factor Surveillance System (BRFSS) data from the US Centers for Disease Control and Prevention in 2011, 2013, and 2015 (most recent years with CVD data)
- CVD endpoints used were myocardial infarction and chronic heart disease
- Use SMOTE[3] algorithm to address imbalance (CVD vs non-CVD) in training data
- Generate predictive models using 9 CVD risk factors (see abstract) and data from 2011 and 2013 to predict BRFSS-reported CVD history
- Test various machine learning algorithms on data unseen during training (2013 and 2015)
- Compare agreement between predictions and reported prevalence data overall and by geography. Use the kappa statistic to account for expected accuracy from random guessing

Results

- Self-reported BRFSS prevalence estimates are in good agreement with estimates from other clinical sources (Figure 1)
- Body mass index (BMI) class (obese) and high cholesterol reveal the largest disparities

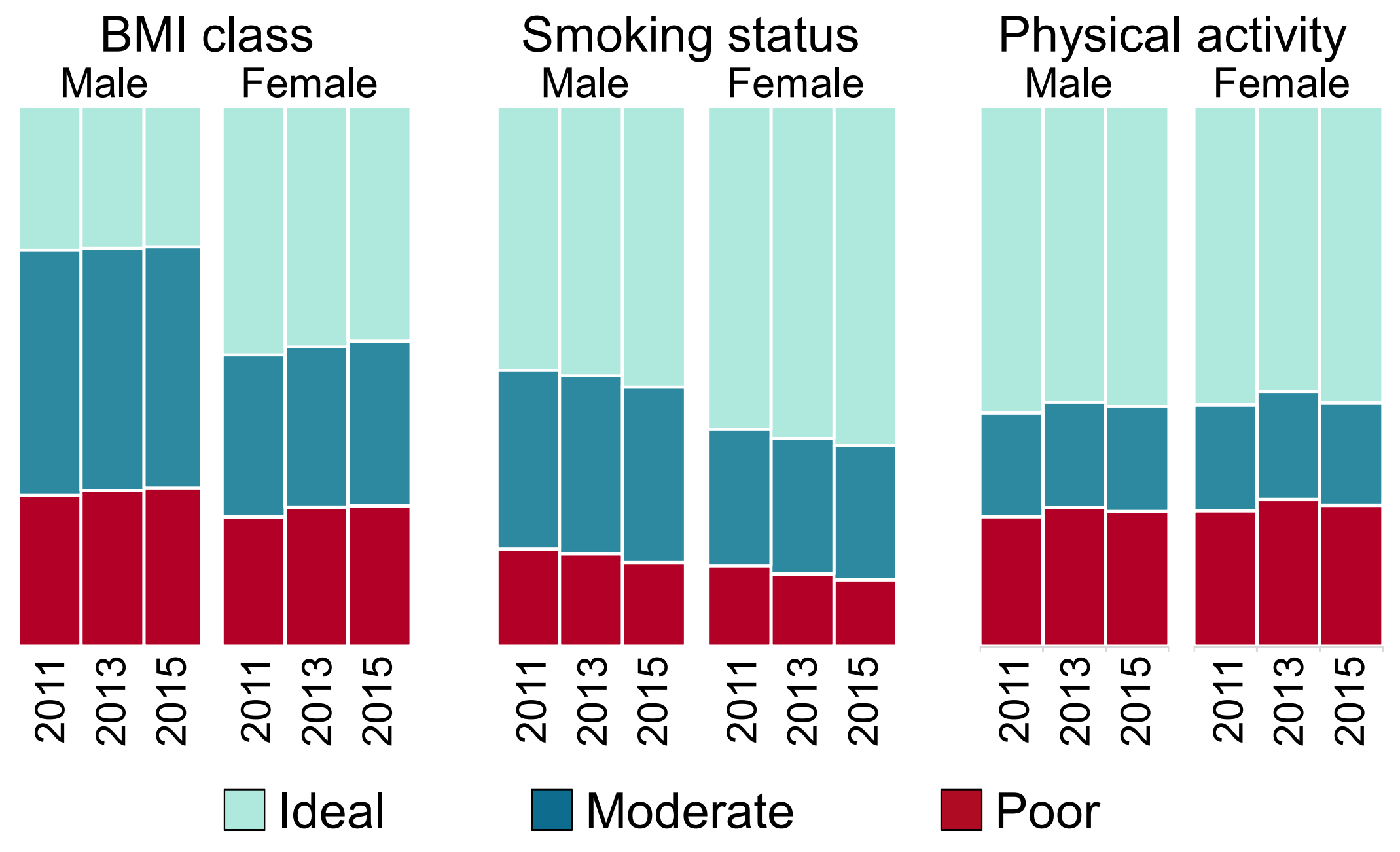
Figure 1: Estimates of CVD risk factor prevalence



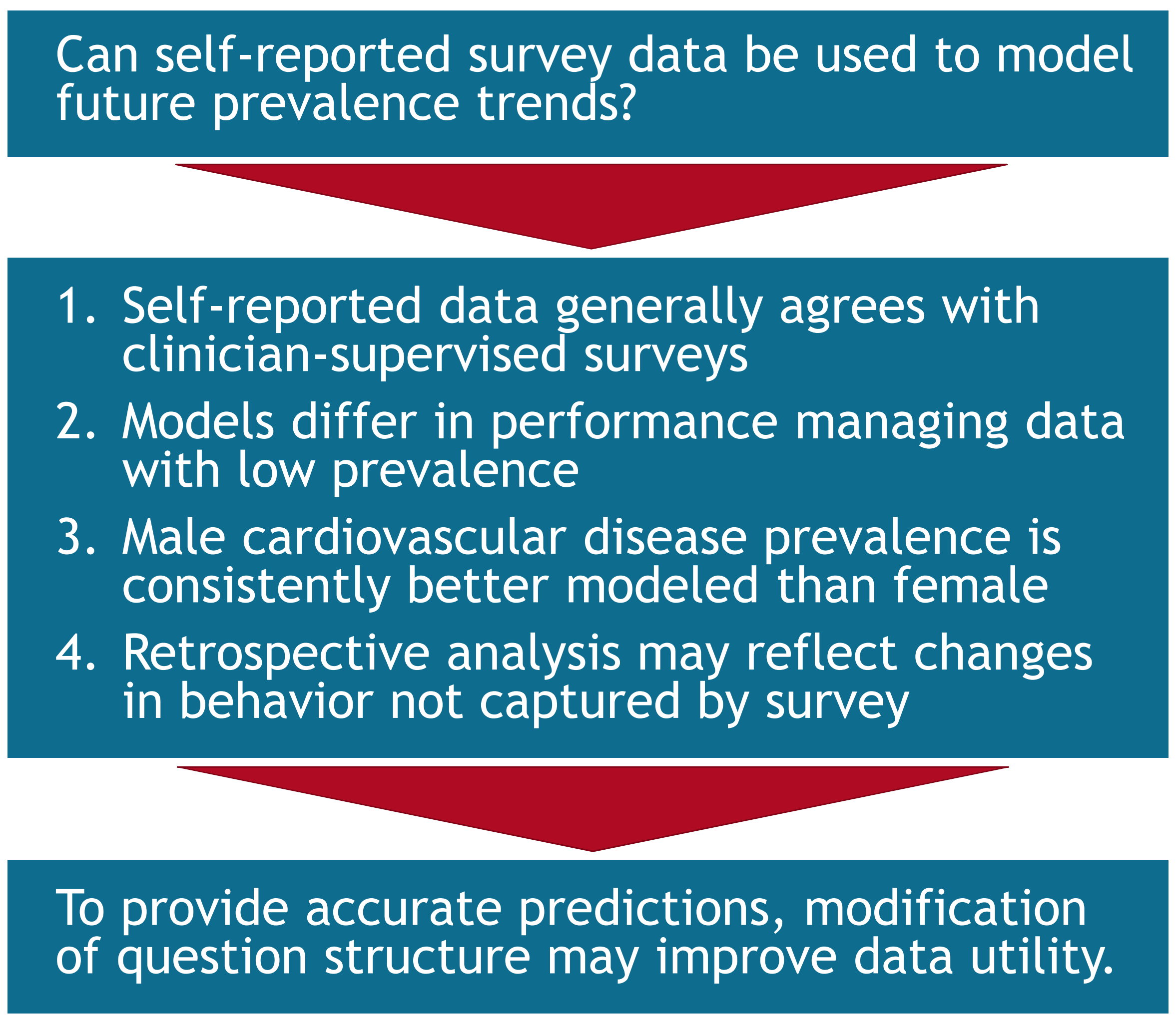
Prevalence of key risk factors for CVD in self-reported BRFSS data (2011 and 2013) and clinically-overseen NHANES (2009-2012), and unpublished NCHS and NHLBI data². BP, Bloos pressure; NHANES, National Health and Nutrition Examination Survey.

- Prevalence of some risk factors reveal short-term trends and differences between males and females (Figure 2)

Figure 2: Gender trends 2011-2015 in select CVD risk factor prevalence



BRFSS prevalence data for three CVD risk factors were stratified by gender across three data sets. Classifications according to ref 2



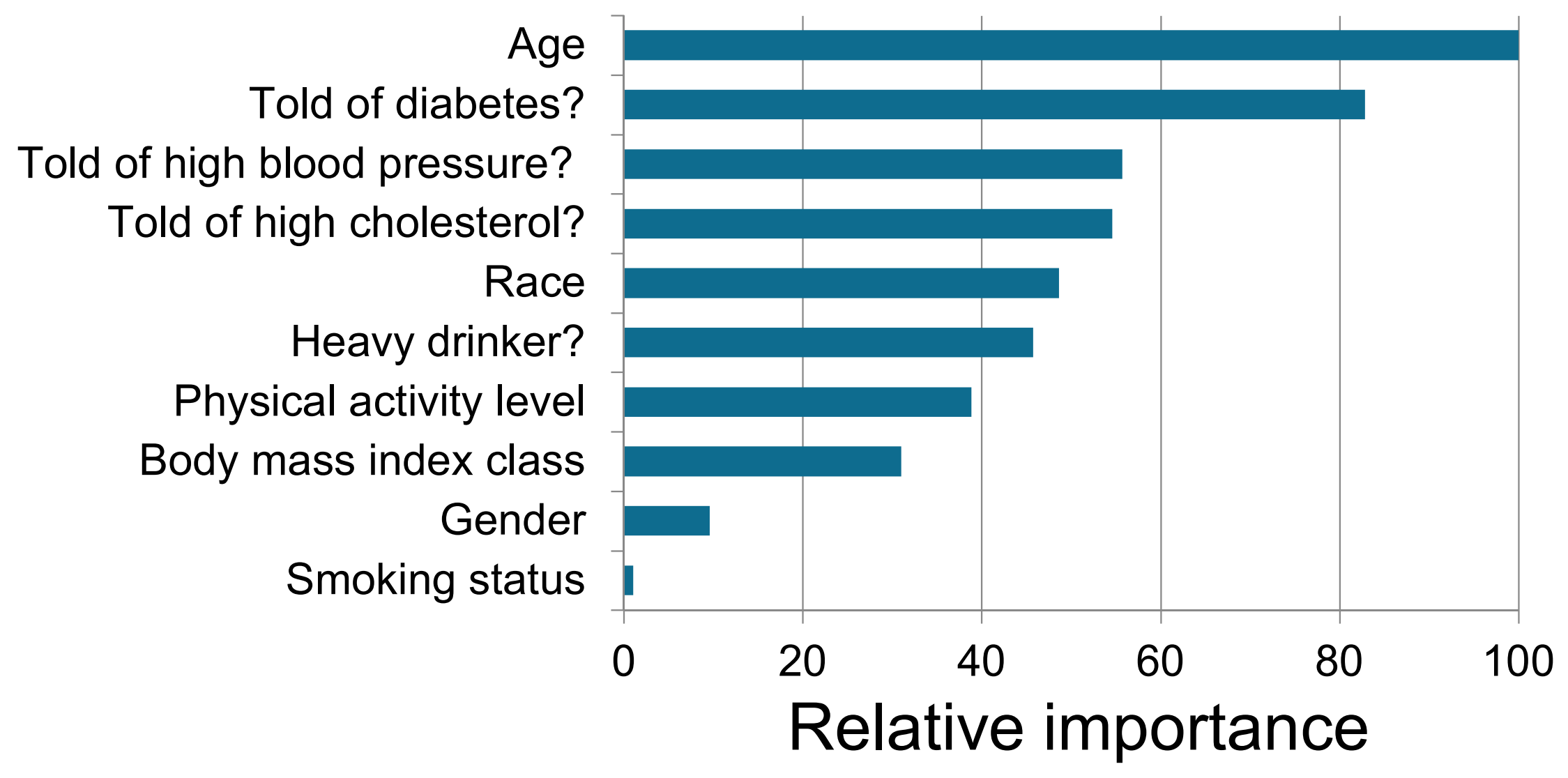
Can self-reported survey data be used to model future prevalence trends?

1. Self-reported data generally agrees with clinician-supervised surveys
2. Models differ in performance managing data with low prevalence
3. Male cardiovascular disease prevalence is consistently better modeled than female
4. Retrospective analysis may reflect changes in behavior not captured by survey

To provide accurate predictions, modification of question structure may improve data utility.

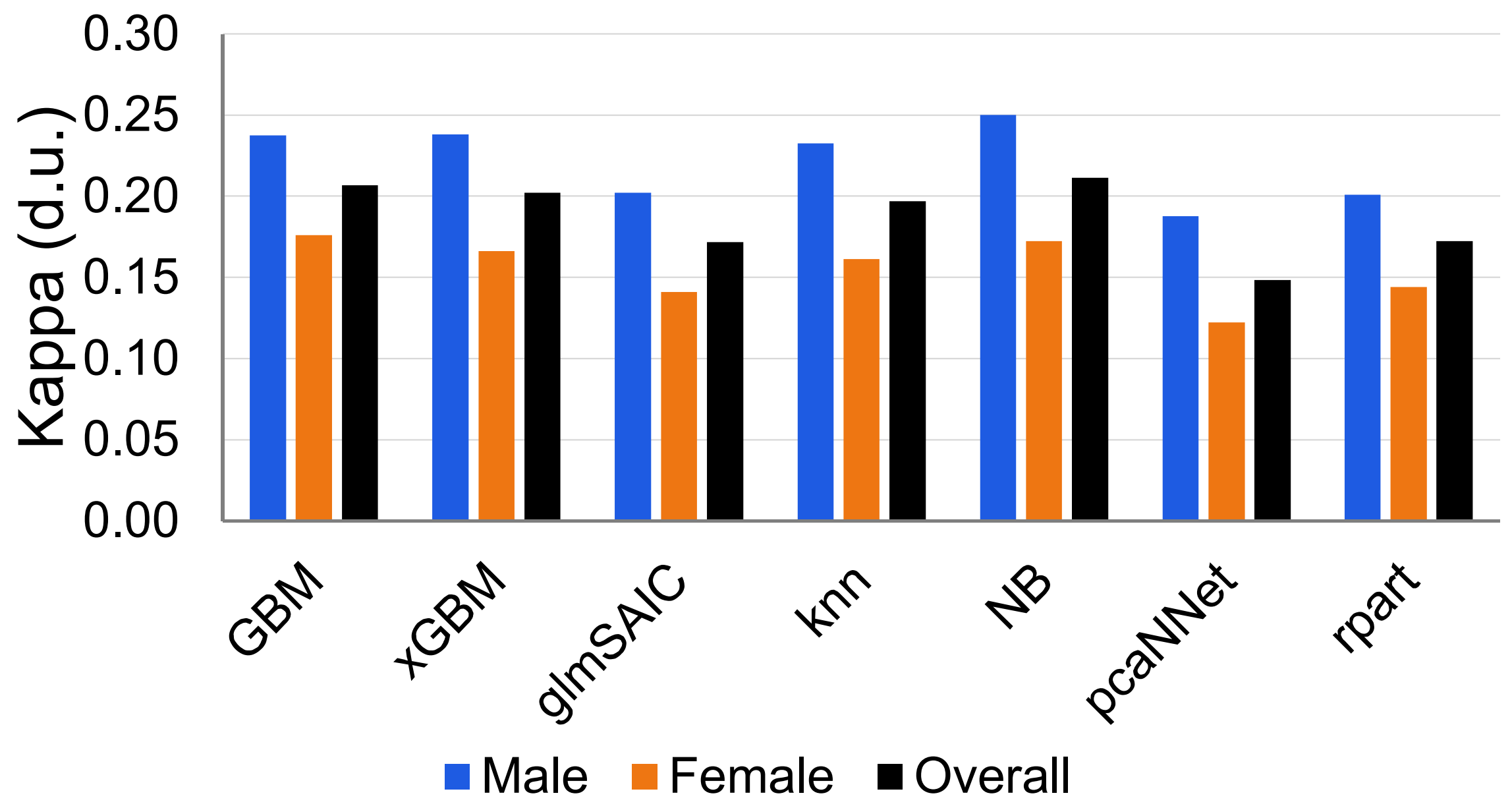
- Ranking of variable importance during model fitting consistently indicates that age, diabetes status, and high blood pressure or cholesterol are key determinants of CVD history (Figure 3)

Figure 3: Average relative variable importance during model fitting



Relative variable importance for CVD risk factors averaged over training results for 7 tree and non-tree based models (see Figure 4) using BRFSS 2011 and 2013 data for model training.

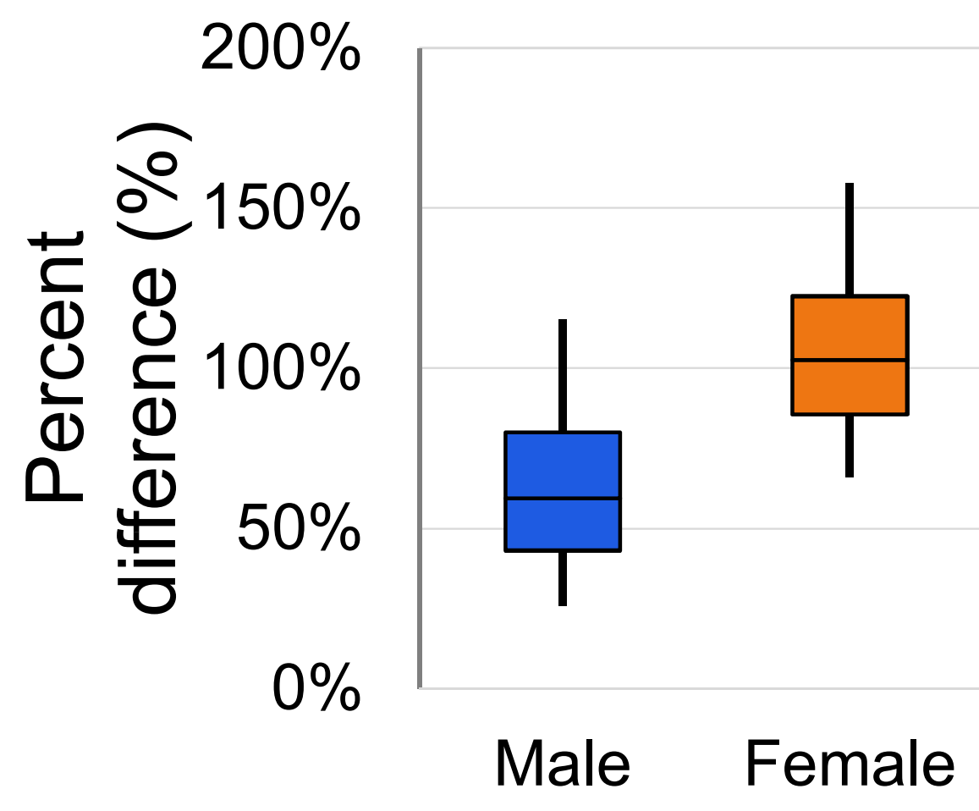
Figure 4: Classification performance of sampled models



Higher kappa values correspond to better classification performance, consistently observed for male data compared to female results. d.u., dimensionless units; GBM, gradient boosting machine; xGBM, extreme gradient boosting machine; glmSAIC, generalized linear model with stepwise feature selection (Akaike information criterion); knn, k-nearest neighbours; NB, Naïve Bayes; pcaNNet, principal component analysis neural network; rpart, recursive partitioning of trees.

- Model performance was assessed using the kappa statistic (Figure 4). The percent difference between predicted and actual prevalence was also determined using the expanded set of models (Figure 5)

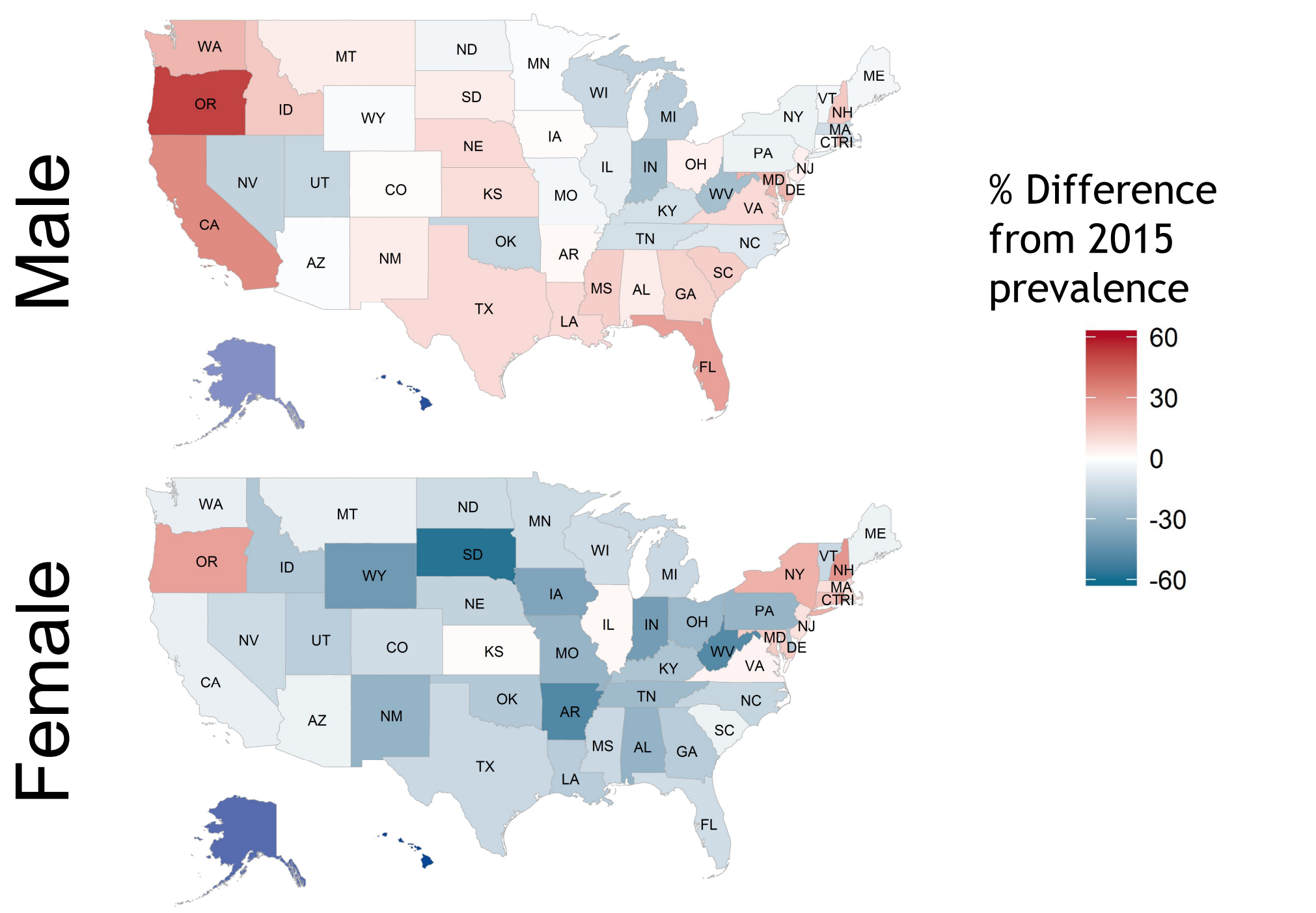
Figure 5: Comparison of all model prediction results with actual prevalence



Geometric means of the percent differences between model predictions and actual prevalence were bootstrapped (10,000 replicates) for male and female data. Whiskers on boxplots correspond to 95% credibility interval.

- For male CVD, model predictions are in better agreement by classification and percent difference than for female data
- Differences were stratified by geography (Figure 6)

Figure 6: Percent difference between Naïve Bayes prediction and actual prevalence data by state in 2015



The range of differences between actual and predicted prevalence for female data is wider than that for males and the nationwide distribution differs.

Discussion

- The greater deviations between model predictions and prevalence data for females may reflect greater changes in behavior
- Poorer predictions for female CVD are in paradox to the mean 20% more training examples available for females than males
- The use of questions without time boundaries in retrospective data collection ("have you ever been told your blood pressure is high?") may confound an accurate snapshot of current disease status
- An expanded set of models tested for performance yields greater variability and may reflect differences in ability to train appropriately on low-prevalence data

References: 1) Pearson-Stuttard et al., Circulation 2016; 133:967-978; 2) Mozaffarian et al., Circulation 2016; 133:e38-e360; 3) Chawla et al., SMOTE: Synthetic Minority Over-sampling Technique. J Artificial Intelligence Res 2002; 16:321-357

PCV94: ISPOR 22nd Annual International Meeting | Boston | May 2017

jason@coreva-scientific.com
+49 (0)76 176 999 422