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

Jason A. Davis, Rhodri Saunders
1. Coreva Scientific, Freiburg, Germany.

objective

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 assess 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, especially on accurate self-reports of cardiovascular disease (CVD) prevalence examined by gender to assess the feasibility of predicting future trends.

Methods: BRFSS data were trimmed to cases for 9 CVD risk factors: gender; age; race; smoking history; physical activity; cholesterol; diabetes; hypertension; and 2013 and 2015 were used to train Bayesian and tree-based algorithms to evaluate predictor performance on unseen data from subsequent years (2013 and 2015) by comparing predictions with reported prevalence.

Results: For algorithms used, predictions of future prevalence were significantly better for males than females (p < 0.001; Sobel’s test for corrections). In the best performing algorithms (Bayesian learners), the mean percent difference from the actual prevalence for males was 3.8±3.5% and females 151±62% (p < 0.05, two-tailed t-test). Data from 2013 yielded better 2-year predictions than 2011. (p < 0.05, 2-tailed t-test). Models trained on the gender combined yielded underestimated prevalence of (p < 0.05). Overall, 2013 BRFSS surveys yielded better accuracy than 2011 for both males and females (p < 0.05, 2-tailed t-test), while for males, there was no significant difference (p = 0.54, two-tailed t-test). Models trained on the gender combined resulted in underestimated prevalence (p < 0.05). Overall, 2013 BRFSS surveys yielded better accuracy than 2011 for both males and females (p < 0.05, 2-tailed t-test).

Discussion: Our findings may reflect more substantial lifestyle changes in females or suggest discussion on changes in how survey 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 than 2011. (p < 0.05, 2-tailed t-test). Models trained on the gender combined yielded underestimated prevalence of (p < 0.05). Overall, 2013 BRFSS surveys yielded better accuracy than 2011 for both males and females (p < 0.05, 2-tailed t-test), while for males, there was no significant difference (p = 0.54, two-tailed t-test). Models trained on the gender combined resulted in underestimated prevalence (p < 0.05).

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

Figure 3: Average relative variable importance during model fitting

Figure 4: Classification performance of sampled models

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

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

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 SMOTE2 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 unsewn 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-overserved NHANES (2009-2012), and unpublished NCHS and NHIGHL datad, 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

Prevalence key risk factors for CVD in self-reported BRFSS data (2011 and 2013) and clinically-overserved NHANES (2009-2012), and unpublished NCHS and NHIGHL datad, 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

- Self-reported data generally agrees with clinician-supervised surveys
- Models differ in performance managing data
- Male cardiovascular disease prevalence is better estimated compared with females
- 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

- Classification performance of sampled models

Figure 4: Classification performance of sampled models

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 actual prevalence 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 more reflective differences in ability to train appropriately on low-prevalence data


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

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