Feature Selection Based on Subpopulations and Propensity Score Matching: A Coronary Artery Disease Use Case using the UK Biobank

Uri Kartoun PhD, Paul Myers PhD, Kristen Severson PhD, Wangzhi Dai PhD,
Kenney Ng PhD, Collin Stultz MD PhD

Unbreak My Heart: Informatics Advances in Cardiovascular Disease / S112 (Nov. 9, 2022)

Uri Kartoun, PhD

Disclosure

I work for IBM Research (Sep. 2016–current).

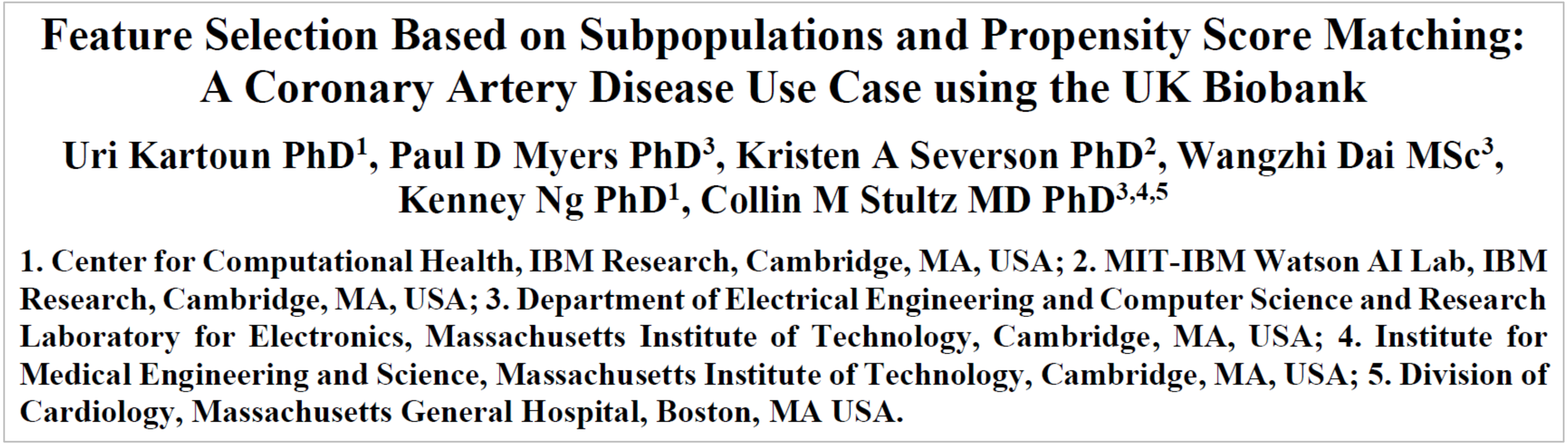
Background

Selecting sub-set of most informative features is crucial in most data-driven scenarios, for example to:

- ❑ Improve efficiency by discarding non-informative features.
- ❑ Minimize the size of the resource required for analyses.
- ❑ Inform data collection design (e.g., for clinical trials).

A new type of a feature selection method

Sub-population-based feature selection



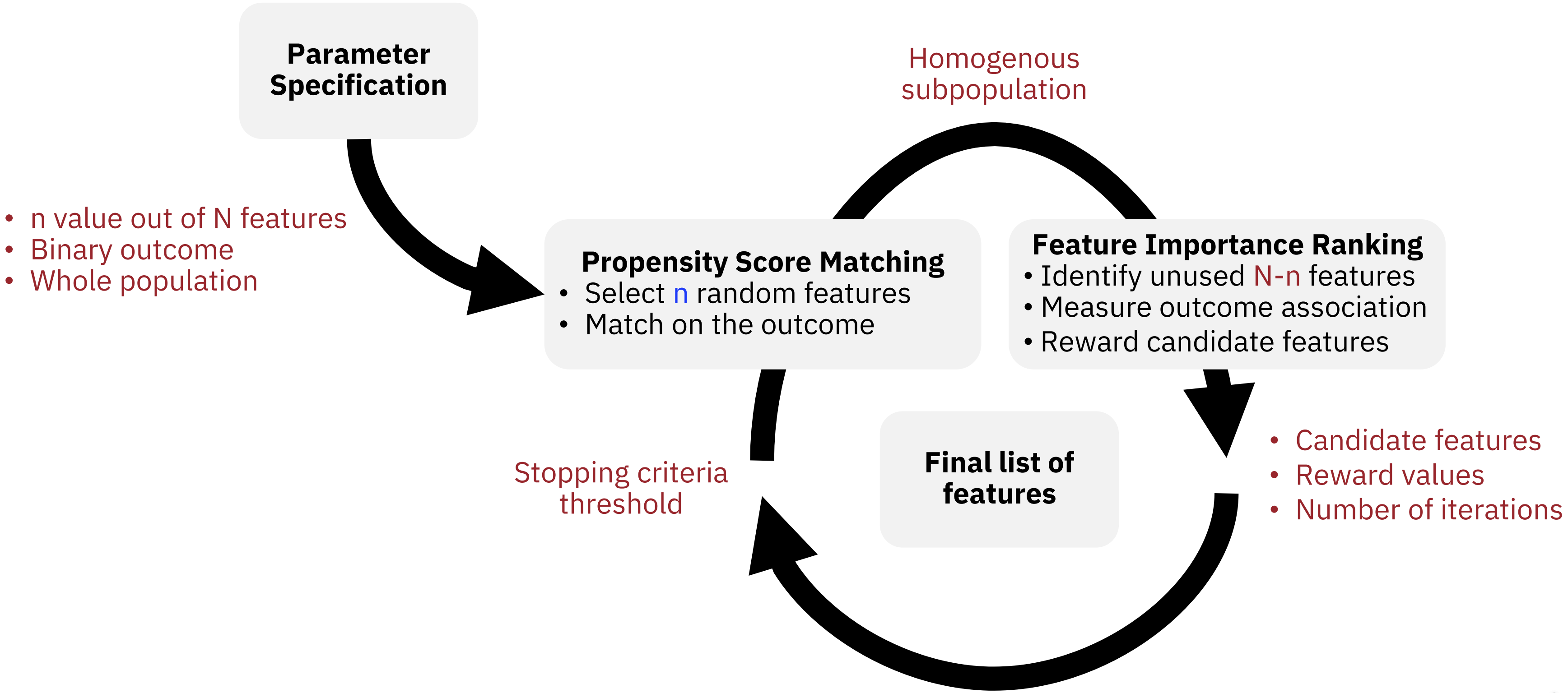
Feature Selection Based on Subpopulations and Propensity Score Matching: A Coronary Artery Disease Use Case using the UK Biobank

Uri Kartoun PhD[1], Paul D Myers PhD[3], Kristen A Severson PhD[2], Wangzhi Dai MSc[3], Kenney Ng PhD[1], Collin M Stultz MD PhD[3,4,5]

1. Center for Computational Health, IBM Research, Cambridge, MA, USA; 2. MIT-IBM Watson AI Lab, IBM Research, Cambridge, MA, USA; 3. Department of Electrical Engineering and Computer Science and Research Laboratory for Electronics, Massachusetts Institute of Technology, Cambridge, MA, USA; 4. Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA, USA; 5. Division of Cardiology, Massachusetts General Hospital, Boston, MA USA.

How does the method work?

Sub-population-based feature selection



Parameter Specification

- n value out of N features
- Binary outcome
- Whole population

Propensity Score Matching

- Select n random features
- Match on the outcome

Homogenous subpopulation

Feature Importance Ranking

- Identify unused N-n features
- Measure outcome association
- Reward candidate features

- Candidate features
- Reward values
- Number of iterations

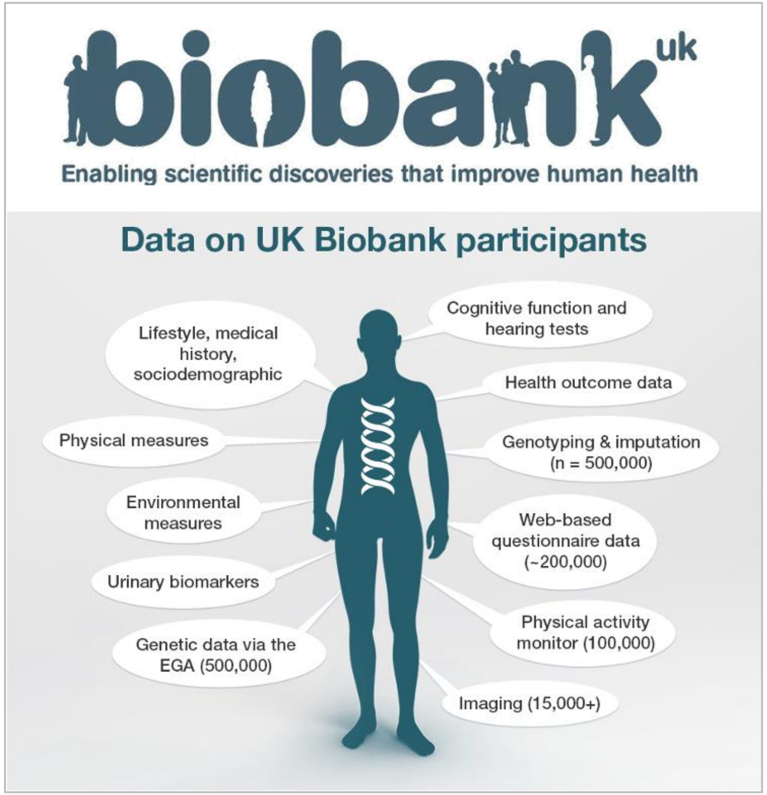
Final list of features

Stopping criteria threshold

What is the UK Biobank?

- ❑ A prospective study with over 500,000 individuals aged 40–69 years recruited through 22 assessment centers in the UK.

- ❑ Questionnaires and physical measures were collected at recruitment, and all participants are followed for outcomes through linkage to national health-related datasets.



biobank uk

Enabling scientific discoveries that improve human health

Data on UK Biobank participants

Lifestyle, medical history, sociodemographic

Physical measures

Environmental measures

Urinary biomarkers

Genetic data via the EGA (500,000)

Cognitive function and hearing tests

Health outcome data

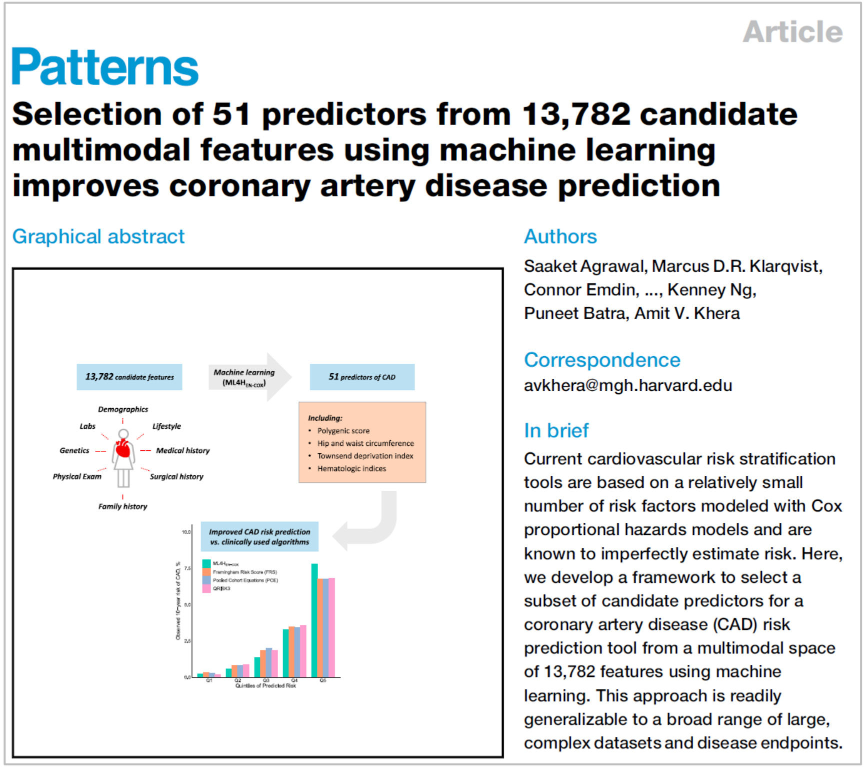
Genotyping & imputation (n = 500,000)

Web-based questionnaire data (~200,000)

Physical activity monitor (100,000)

Imaging (15,000+)

A comparison with leading methods



Article

Patterns

Selection of 51 predictors from 13,782 candidate multimodal features using machine learning improves coronary artery disease prediction

Graphical abstract

13,782 candidate features
Machine learning (ML4H$_{EN-COX}$)
51 predictors of CAD
Demographics
Labs
Lifestyle
Genetics
Medical history
Physical Exam
Surgical history
Family history
Including:
• Polygenic score
• Hip and waist circumference
• Townsend deprivation index
• Hematologic indices
Improved CAD risk prediction vs. clinically used algorithms
ML4H$_{EN-COX}$
Framingham Risk Score (FRS)
Pooled Cohort Equations (PCE)
QRISK3
Observed 10−year risk of CAD, %
0.0
2.5
5.0
7.5
10.0
Q1
Q2
Q3
Q4
Q5
Quintiles of Predicted Risk

Authors

Saaket Agrawal, Marcus D.R. Klarqvist, Connor Emdin, ..., Kenney Ng, Puneet Batra, Amit V. Khera

Correspondence

avkhera@mgh.harvard.edu

In brief

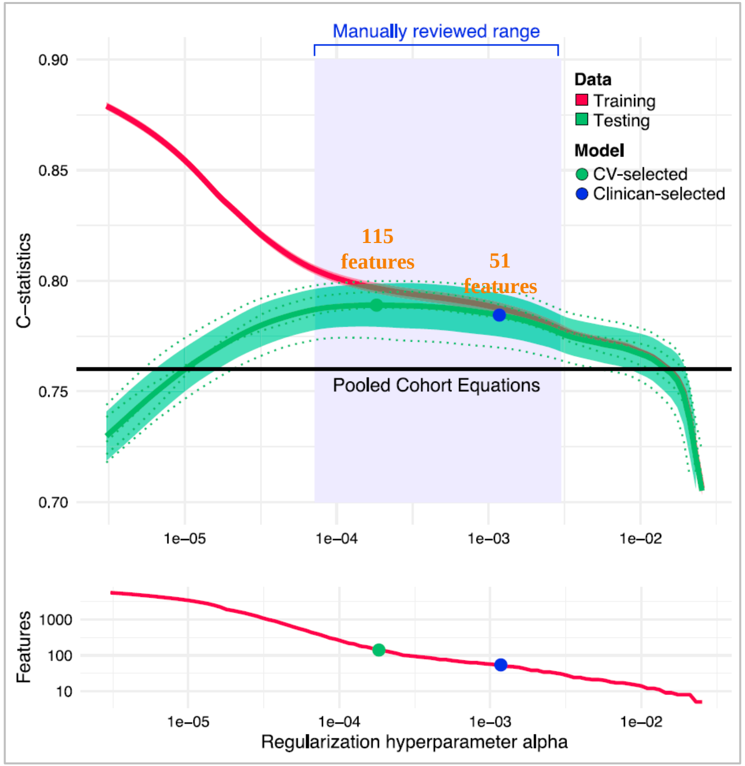
Current cardiovascular risk stratification tools are based on a relatively small number of risk factors modeled with Cox proportional hazards models and are known to imperfectly estimate risk. Here, we develop a framework to select a subset of candidate predictors for a coronary artery disease (CAD) risk prediction tool from a multimodal space of 13,782 features using machine learning. This approach is readily generalizable to a broad range of large, complex datasets and disease endpoints.

A novel feature selection method

Machine Learning for Health—Elastic Net regularized Cox model (ML4HEN-COX)

- **A two-step human-in-the-loop approach**
  - **Parameter optimization**
  - **Clinician review to refine features**

A comparison with leading methods



Performance is bounded

115 features

CI = 0.797 (0.784–0.810)

Use case using the UK Biobank

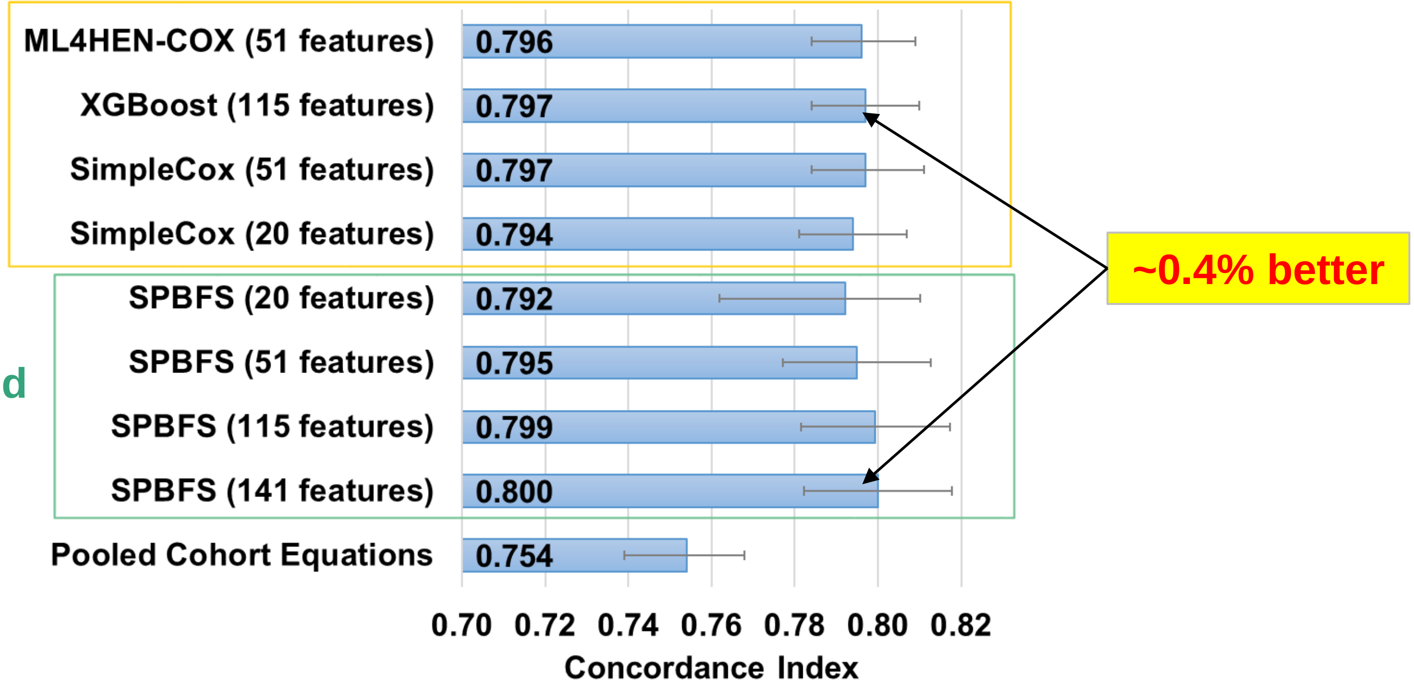
- ❑ 173,274 patients
    - ❑ Development set (N = 138,619)
    - ❑ Holdout set (N = 34,655)
- ❑ 13,782 features
    - ❑ Comorbidities, surgical history, labs, medications, demographics, family history, genetics
- ❑ A binary outcome
    - ❑ 10-year incident of coronary artery disease

A comparison with leading methods

Holdout set (N = 34,655)

Leading feature selection methods

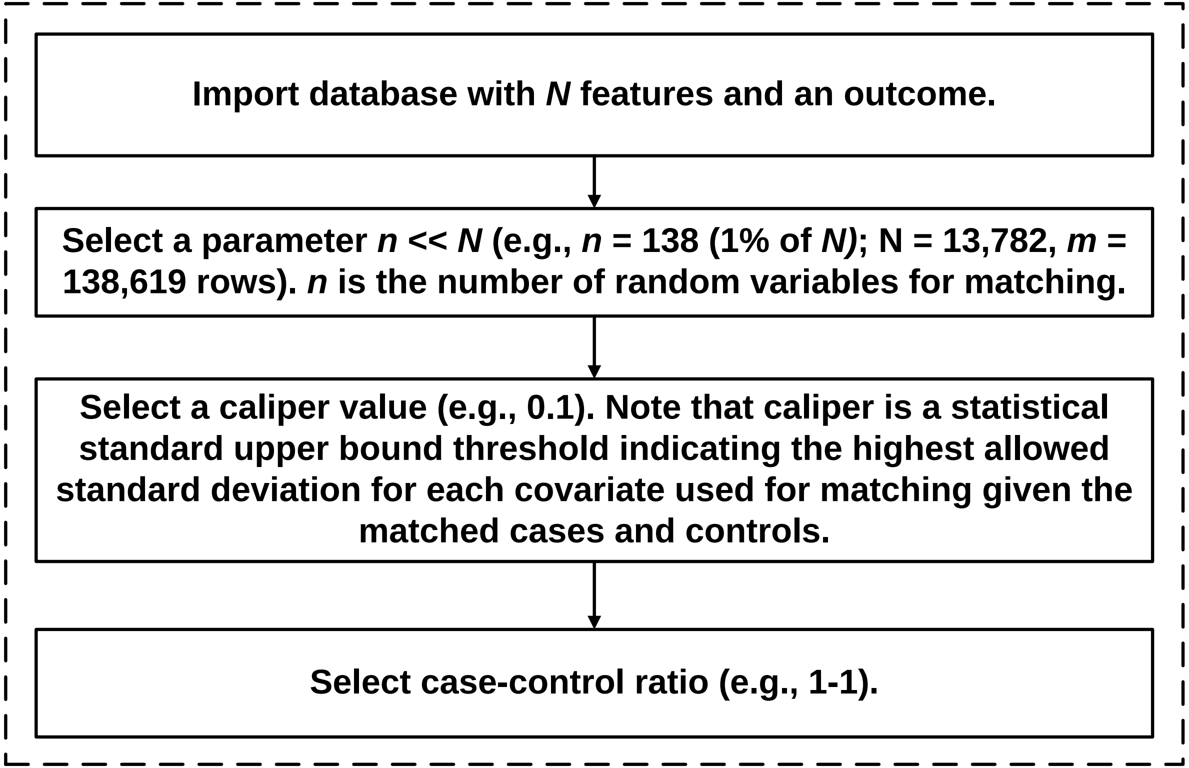
Sub-population-based feature selection



~0.4% better

How does the method work?

Step 1



Import database with *N* features and an outcome.

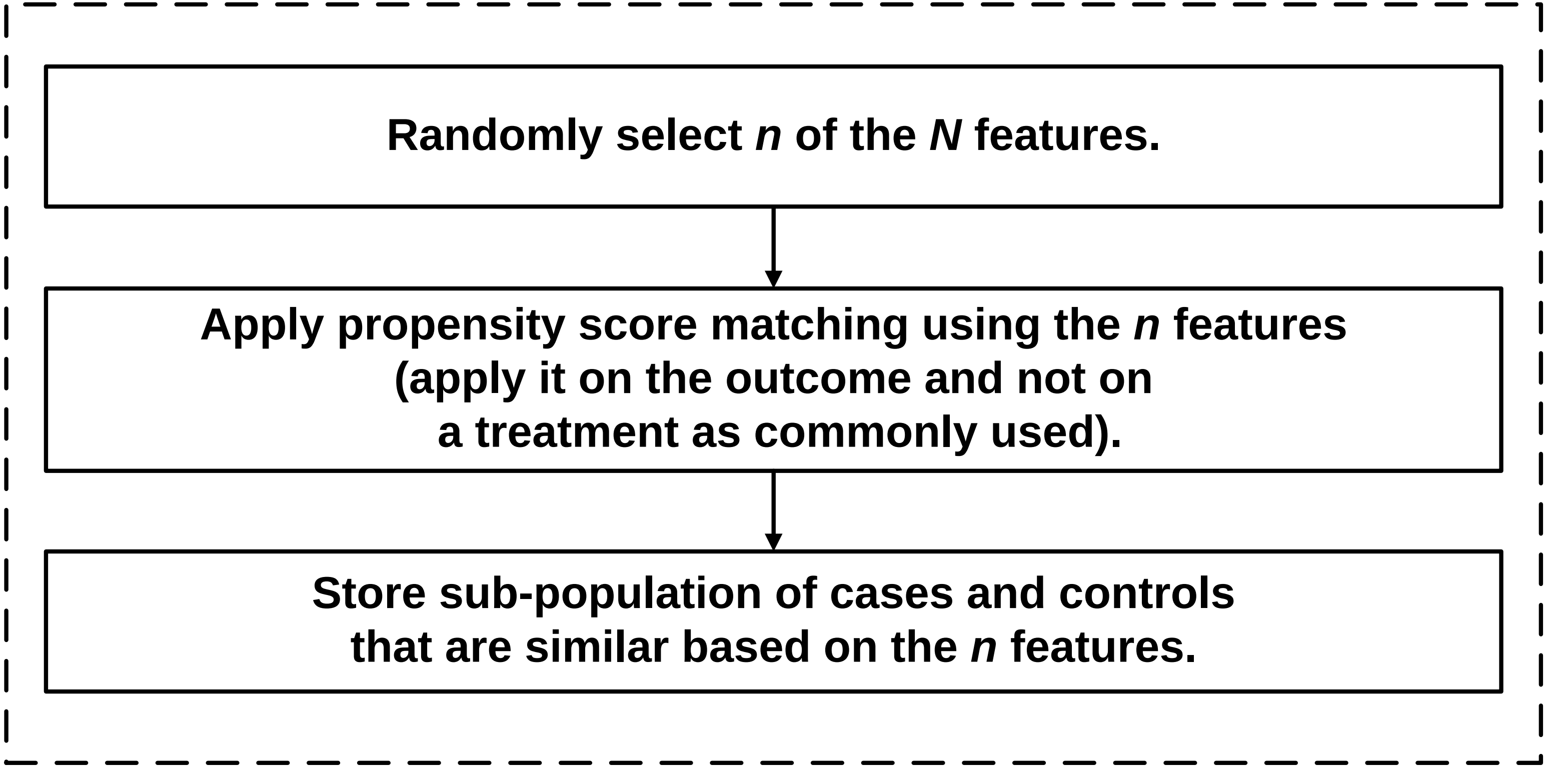
Select a parameter $n \ll N$ (e.g., n = 138 (1% of N); N = 13,782, m = 138,619 rows). n is the number of random variables for matching.

Select a caliper value (e.g., 0.1). Note that caliper is a statistical standard upper bound threshold indicating the highest allowed standard deviation for each covariate used for matching given the matched cases and controls.

Select case-control ratio (e.g., 1-1).

How does the method work?

Step 2



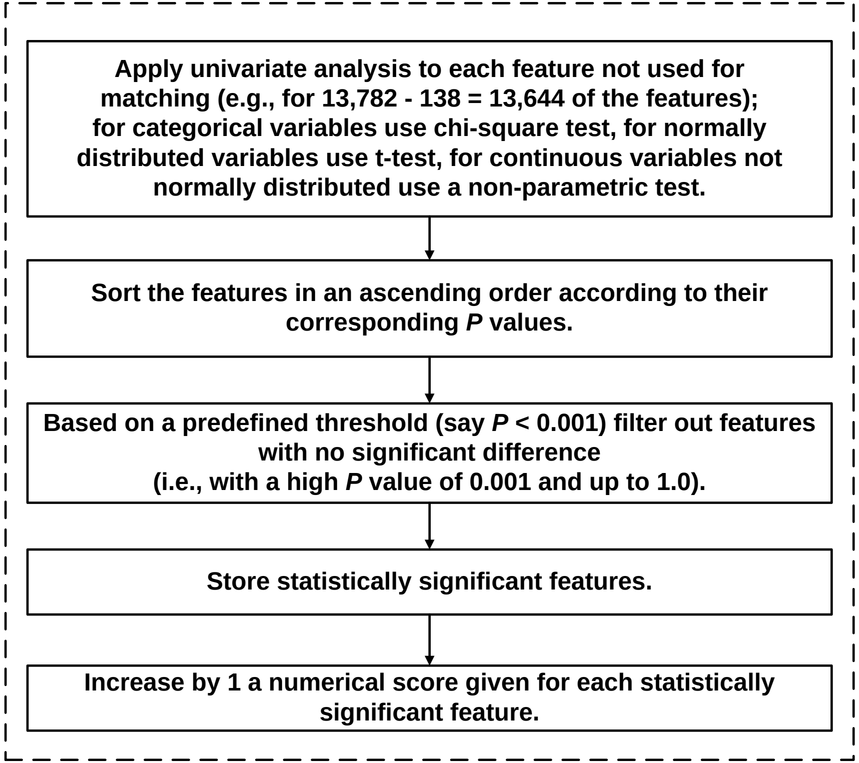
Randomly select *n* of the *N* features.

Apply propensity score matching using the *n* features (apply it on the outcome and not on a treatment as commonly used).

Store sub-population of cases and controls that are similar based on the *n* features.

How does the method work?

Step 3



Apply univariate analysis to each feature not used for matching (e.g., for 13,782 - 138 = 13,644 of the features); for categorical variables use chi-square test, for normally distributed variables use t-test, for continuous variables not normally distributed use a non-parametric test.

Sort the features in an ascending order according to their corresponding *P* values.

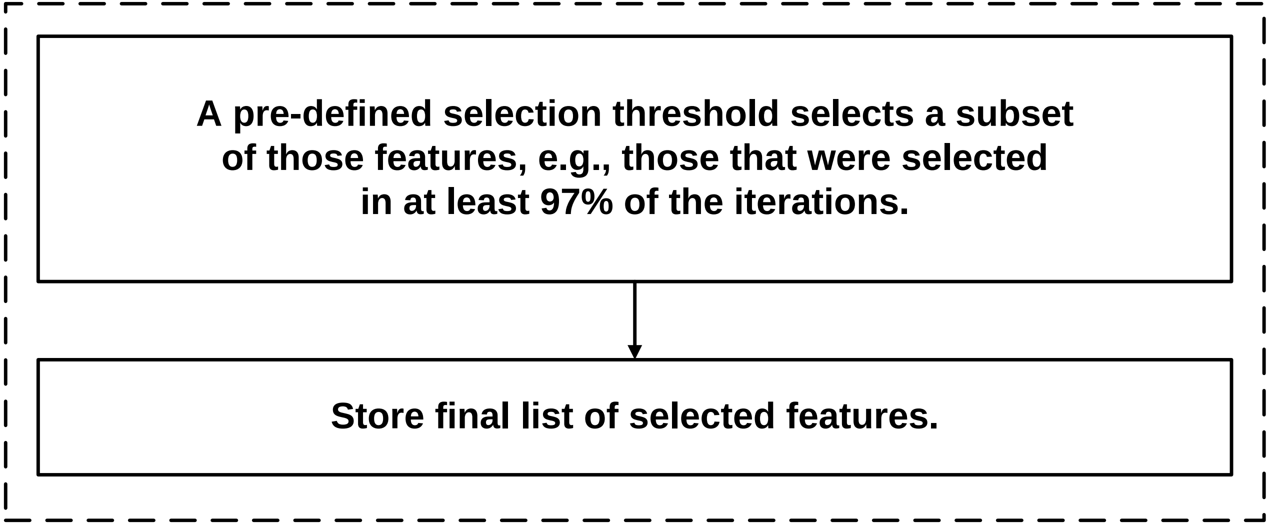
Based on a predefined threshold (say *P* < 0.001) filter out features with no significant difference (i.e., with a high *P* value of 0.001 and up to 1.0).

Store statistically significant features.

Increase by 1 a numerical score given for each statistically significant feature.

How does the method work?

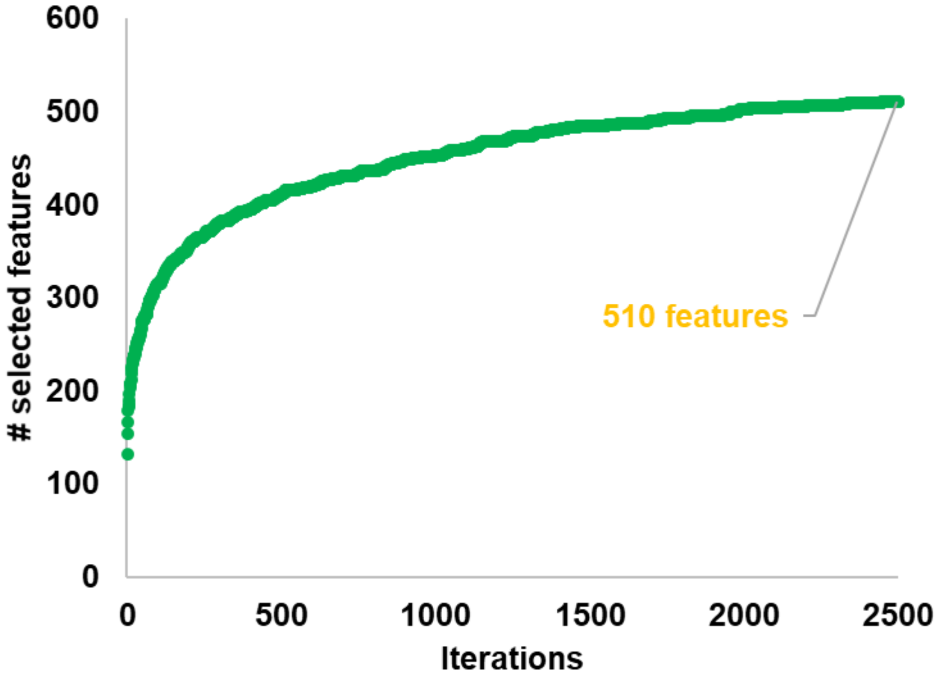
Step 4



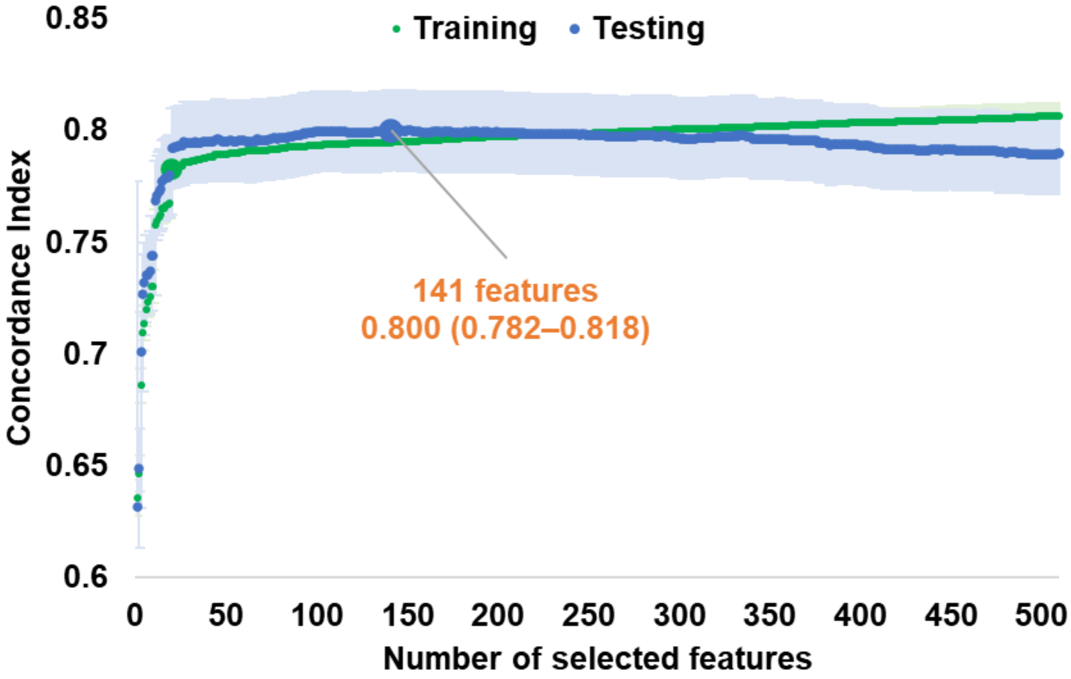
A pre-defined selection threshold selects a subset of those features, e.g., those that were selected in at least 97% of the iterations.

Store final list of selected features.

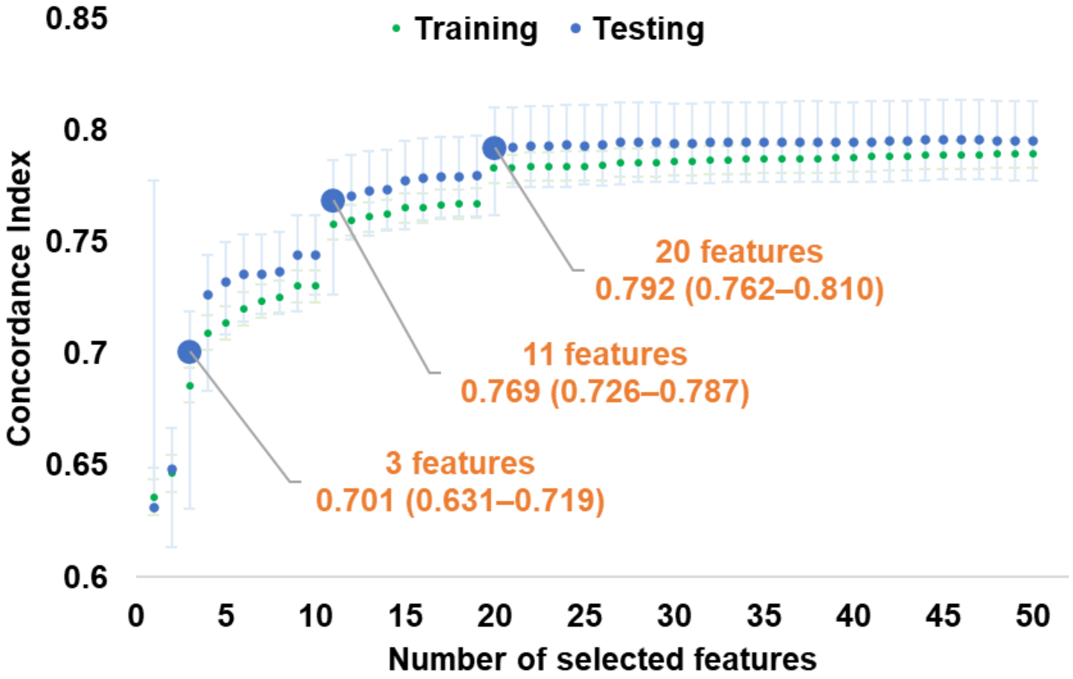
2,500 Iterations



Performance Evaluation



Performance Evaluation



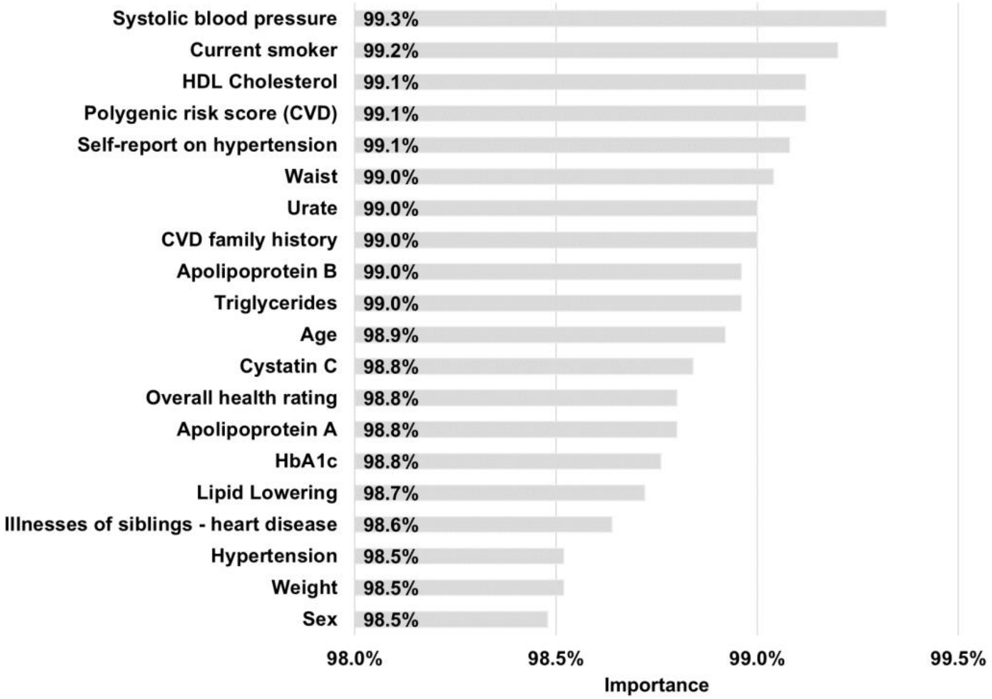
20 features
0.792 (0.762–0.810)

11 features
0.769 (0.726–0.787)

3 features
0.701 (0.631–0.719)

20 Selected Features

(All were rewarded in >98.5% of the iterations; >2,462 out of 2,500)



20 Selected Features

(All were rewarded in >98.5% of the iterations; >2,462 out of 2,500)

Demographic / Behavioral
Age
Current smoker
Sex

Physical
Weight
Waist

Labs
HbA1c
HDL Cholesterol
Apolipoprotein A
Apolipoprotein B
Cystatin C
Triglycerides
Urate

Comorbidities
Hypertension

Vitals
Systolic blood pressure

Family History
CVD family history

Drugs
Lipid Lowering

Other
Polygenic risk score (CVD)
Overall health rating
Self-report on hypertension
Illnesses of siblings - heart disease

20 Selected Features

(All were rewarded in >98.5% of the iterations; >2,462 out of 2,500)

Components of the Pooled Cohort Equations

2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk

Demographic / Behavioral
Age
Current smoker
Sex

Physical
Weight
Waist

Labs
HbA1c
HDL Cholesterol
Apolipoprotein A
Apolipoprotein B
Cystatin C
Triglycerides
Urate

Comorbidities
Hypertension

Vitals
Systolic blood pressure

Family History
CVD family history

Drugs
Lipid Lowering

Other
Polygenic risk score (CVD)
Overall health rating
Self-report on hypertension
Illnesses of siblings - heart disease

Next steps

❑ Hyperparameter optimization

❑ Convergence assessment
- ❑ How many iterations? 100? 2,500? 10,000? 1M? Other?
- ❑ How well does the method do within a small number of iterations (e.g., 10)?

❑ Find use cases
- ❑ At IBM
- ❑ Externally

❑ Help others to use our publicly available R package
- ❑ Simple to install and use: www.github.com/IBM/spbfs
  - ❑ install.packages("devtools"); library(devtools)
  - ❑ install_github("IBM/spbfs"); library('spbfs')

Conclusions

Sub-population-based feature selection

- ❑ We developed a new type of feature selection method incorporating propensity matching applied iteratively to subpopulations.

- ❑ Our method holds advantages
  - ❑ Comparable prediction performance to leading methods
    - ❑ A comparable performance using a small number of features.
    - ❑ A 0.4% performance boost with a large number of features.
      - ❑ Performance boost may be higher with additional iterations / tuning.
  - ❑ No need for manual review.
  - ❑ Publicly available as an R package.

Thank you!

uri.kartoun@ibm.com

www.github.com/IBM/spbfs