**PROBAST**

Study:

A Machine Learning Approach to Classifying Self-Reported Health Status in a Cohort of Patients with Heart Disease Using Activity Tracker Data

Step 2: Type of prediction study

**Is the study a diagnostic or a prognostic study?**

Diagnostic

**Is the study a development only, development and validation or validation only study?**

Development only

**What is the model of interest?**

Random forest

**What is the outcome of interest?**

Estimate patient wellbeing

Step 3: Assess risk of bias

**Domain 1: Participants**

**Describe the sources of data and criteria for participant selection**

A set of 200 patients with SIHD were recruited for a feasibility study conducted by Cedars-Sinai Medical Center from 2017to 2018 to predict surrogate markers of major adverse cardiac events (MACE), including myocardial infarction, arrhythmia, and hospitalization due to heart failure, using biometrics, wearable sensors, patient-reported surveys, and other biochemical markers.

**1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?**

**Y**

**1.2 Were all inclusions and exclusions of participants appropriate?**

Y

**Risk of bias introduced by selection of participants:**

Low

**Rationale of bias rating**

No further exclusion criteria given

**Domain 2: Predictors**

**List and describe predictors included in the final model, e.g. definition and timing of assessment**

Fitbit data: steps, distance, activity, heart rate

**2.1 Were predictors defined and assessed in a similar way for all participants?**

Y

**2.2 Were predictor assessments made without knowledge of outcome data?**

Y

**2.3 Are all predictors available at the time the model intended to be used?**

Y

**Risk of bias introduced by predictors or their assessment**

Low

**Rationale of bias rating**

Wearable data are applicable and independent

**Domain 3: Outcome**

**Describe the outcome, how it was defined and determined, and the time interval between predictor assessment and outcome determination:**

This analysis uses data from eight PROMIS instruments: Global Physical Health and Global Mental Health, which are two composite scores from the Global-10 short form [35]; Fatigue-Short Form 4a; Physical Function-Short Form 10a; Emotional Distress-Anxiety-Short Form 6a;Depression-Short Form 4a; Social Isolation-Short Form 4a; and Sleep Disturbance-Short Form 4a. Each questionnaire either asks about current health or has a recall period of the previous seven days, so they are appropriate for weekly administration. The T metric method was used to standardize scores for each type to a mean of 50 and a standard deviation of 10, with a range between 0 and 100 [15], [36]. Symptom (i.e., Fatigue, Anxiety, Depression, Social Isolation, and Sleep Disturbance)scores of 60 or higher are one standard deviation above the average, which is defined as moderate to severe symptom severity. For function (i.e., Global Physical Health, Global Mental Health, and Physical Function), scores less than 40 are classified as moderate to severe, meaning less functional ability than normal. For this study, PRO scores were predicted in two ways: regression was used to predict PRO scores from patient activity tracker data, and classification was used to determine whether subjects’ PRO scores were above the threshold for at least mod-erate severity.

**3.1 Was the outcome determined appropriately?**

Y

**3.2 Was a pre-specified or standard outcome definition used?**

Y

**3.3 Were predictors excluded from the outcome definition?**

Y

**3.4 Was the outcome defined and determined in a similar way for all participants?**

Y

**3.5 Was the outcome determined without knowledge of predictor information?**

PN

**3.6 Was the time interval between predictor assessment and outcome determination appropriate?**

Y

**Risk of bias introduced by the outcome or its determination**

Low

**Rationale of bias rating**

Standard patient wellbeing questionnaires were taken. Patients may have used information from Fitbit to fill in questionnaires however this may still be a appropriate.

**Domain 4: Analysis**

**Describe number of participants, number of candidate predictors, outcome events and events per candidate predictor**

After adopting this data preprocessing approach and using the classification criteria above, a total number of 182 subjects with a total of 1,640 weeks were collected.

**Describe how the model was developed, predictor selection and risk group definition**

The features for each of the seven days were appended into a single feature vector, which was then used as the input for binary classification of each PRO score. Ensemble methods like Adaboost, GBRT (gradient boosting regression tree) and Random Forest (RF) are relatively robust over unbalanced dataset and is capable of generating better classification accuracy than other types of machine learning algorithms [37].

**Describe whether and how the model was validated, either internally (cross validation, random split sample) or externally (e.g. temporal validation, geographical validation, different setting, different type of participants)**

Each of these methods was applied to the dataset using ten-fold cross-validation across subjects in conjunction with grid search to find the optimal parameters for each model.

**Describe the performance measures of the model, e.g. calibration, discrimination, classification, net benefit, and whether they were adjusted for optimism**

**AUC**

**Describe any participants who were excluded from the analysis**

200 to 182. 18 patients removed due to incomplete data

**Describe missing data on predictors and outcomes as well as methods used for missing data**

If data was available for at least four days in a week, missing values were permuted by using the average value of the rest of the week for steps or resting heart rate. Weeks with missing survey scores, as well as those without step and resting heart rate data for more than three days, were removed from the analysis.

**4.1 Were there a reasonable number of participants with the outcome?**

PY

**4.2 Were continuous and categorical predictors handled appropriately?**

Y

**4.3 Were all enrolled participants included in the analysis?**

N

**4.4 Were participants with missing data handled appropriately?**

Y

**4.5 Was selection of predictors based on univariable analysis avoided?**

Y

**4.6 Were complexities in the data (e.g. censoring, competing risks, sampling of controls)**

**accounted for appropriately?**

Y

**4.7 Were relevant model performance measures evaluated appropriately?**

PY

**4.8 Were model overfitting and optimism in model performance accounted for?**

Y

**4.9 Do predictors and their assigned weights in the final model correspond to the results**

**from multivariable analysis?**

Y

**Risk of bias introduced by the analysis**

Low

**Rationale of bias rating**

Appropriate handling of missing data and good validation approach. There could be more metrics but AUC is a good overall metric. Reasonable amount of outcomes.

**Overall Risk of bias**

Low