**PROBAST**

Study:

Machine Learning Approaches to Classify Self-Reported Rheumatoid Arthritis Health Scores Using Activity Tracker Data:Longitudinal Observational Study

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?**

Hidden Markov model

**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 cohort of 470 eligible patients with RA was recruited by researchers at the Global Healthy Living Foundation and the University of Alabama at Birmingham, and 278 (59.1%)of them qualified for participation in the main study after successfully meeting adherence thresholds during an initial2-week lead-in period.

**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 eligibility criteria given

**Domain 2: Predictors**

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

Data from wearable device: step count, sleep data, calories burned, activity

**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**

Applicable and independent and assessed equally

**Domain 3: Outcome**

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

In this study, patients were asked to actively input information regarding 6 PROMIS scores on a weekly basis through several questionnaires on the study-specific ArthritisPower app. “Symptom” (pain interference, fatigue, and sleep disturbance)scores of 60 (1 SD above the average of 50) or higher were defined as moderate to severe symptom severity [31]. Similarly,“ function” (physical function and social activity) scores of 40 (1 SD below the average of 50) or below were defined as moderate to severe symptom severity, meaning less functional ability than normal. In this study, multiclass classification techniques were used to classify PRO state or score transitions over time, and binary classification techniques were used to determine whether patients’ PRO scores were above or below the critical threshold for at least moderate symptom or functional severity.

**3.1 Was the outcome determined appropriately?**

PY

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

PN

**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**

High

**Rationale of bias rating**

Patients could have used information from their wearable (e.g. step count) in their PROs. Also binary features were made from continuous features. Thresholds could have been made during analysis and it leads to loss of information.

**Domain 4: Analysis**

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

219 patients ended up remaining in the data set.

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

the RF model was used to establish a baseline performance that the HMM could potentially improve on. Through hyperparameter tuning, we determined that each RF instance should consist of 100 estimators (decision trees), use the Gini Index as criteria for splitting, allow a maximum depth of 25 to prevent overfitting, and require at least 10 training samples as the minimum threshold for splitting.

Hidden Markov model.

**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)**

**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**

Based on these specifications, 59 of the 278 originally qualifying patients were excluded, and thus, 219 patients ended up remaining in the data set

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

It was first decided that the criteria for dropping a given week of data prior to model training and evaluation would be if there were at least3 days of missing data for any of the 15 Fitbit-derived features. We excluded patients from our data set who had more than 2weeks of dropped rows at any time during the 12-week study period.

In addition, prior to training our machine learning models to generate predictions for each of the PRO scores, we imputed remaining missing feature data based on the corresponding feature means from the previous week.

**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?**

PN

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

N

**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**

High

**Rationale of bias rating**

Amount of outcomes not described, however there are probably enough outcomes with 219 patients and weekly outcomes. A considerable amount of patients were removed due to missing data. Only AUC score reported, other metrics such as sensitivity may have been useful. Validation approach not described.

**Overall Risk of bias**

High