**PROBAST**

Study:

Classifying Self-Reported Rheumatoid Arthritis Flares Using Daily Patient-Generated Data From a Smartphone App: Exploratory Analysis Applying Machine Learning Approaches

Step 2: Type of prediction study

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

Prognostic

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

Development only

**What is the model of interest?**

Logistic regression

**What is the outcome of interest?**

Future self-reported flares

Step 3: Assess risk of bias

**Domain 1: Participants**

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

This study was a post hoc analysis of data from the first phase of the Remote Monitoring of Rheumatoid Arthritis (REMORA)study [12], which involved 20 patients with RA using a smartphone app to track their daily symptoms over 3 months.

Eligibility criteria were(1) clinician-verified RA, (2) treated at a specific outpatientclinic, (3) willingness to participate, and (4) able to providewritten consent.

**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**

Reasonable eligibility criteria given

**Domain 2: Predictors**

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

Participants received prompts every evening to report several symptoms on a 0-10 numerical rating scale based on the RAID scale adapted for daily use [13]: pain, function (“difficulty in doing daily activities”), fatigue (attributed to RA), sleep quality, overall physical and emotional well-being, and ability to cope. Users reported the duration of morning stiffness daily using 1of 7 time intervals. Weekly questionnaires asked patients about self-assessed tender and swollen joint counts and the binary flare question: “Have you experienced a flare in the last week? ” These questions were prompted by a notification every 7 days to complete the weekly question set.

The 7 days up to and including each flare report were treated as the exposure period. For each exposure period, the following5 symptom summary features were calculated for each of the 8daily symptoms: minimum, maximum, mean score, SD, and slope. Isolated daily reports (those not followed by a flare report in the next 6 days) were discarded, so every remaining exposure period contained at least 2 daily data points. Although not prompted, participants were able to answer the weekly flare question at any time during the week outside of the 7-dayschedule, resulting in some partially overlapping exposure periods.

The patient-reported symptom scores were collected using integer numerical rating scales from 0 to 10 (morning stiffness on a 7-point ordinal scale). For this exploratory analysis, all symptoms were treated as continuous variables. This approach was chosen because it allows ease of comparison with other work in intraindividual pain variability.

**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**

Predictors were consistently defined and assessed across participants.

**Domain 3: Outcome**

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

We treated each weekly flare report as a binary outcome. It was left up to the patient to decide what was classified as a flare.

Although not prompted, participants were able to answer the weekly flare question at any time during the week outside of the 7-dayschedule, resulting in some partially overlapping exposure periods. In that case, we allowed the intersecting daily symptom reports to correspond to multiple outcomes. Where the same participant responded more than once on the same date, we assumed later-recorded responses superseded earlier ones

**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?**

PN

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

Y

**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**

Due to overlapping outcomes and patient self reports outcomes may not be consistent and comparable between patients. However, this is normal for PROs. Still low to moderate risk.

**Domain 4: Analysis**

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

The collected data set comprised 20 unique participants completing a total of 1325 daily and 213 weekly questionnaires.

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

We fitted 3 distinct classes of binary classification models to the data: logistic regression with and without elastic net regularization, a random forest, and naive Bayes.

**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)**

A 10-fold cross-validation was performed, with 18 (90%) participants comprising the training sets and the remaining 2 (10%) participants comprising the test sets. The validation was repeated 10 times, each time reserving2 different participants for testing.

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

AUC, SEN, SPE, PRE, REC

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

None

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

Weeks with an unanswered (missing) flare question are not included in this analysis.

no imputation was performed on missing values.

**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?**

Y

**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?**

Y

**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?**

U

**Risk of bias introduced by the analysis**

Low

**Rationale of bias rating**

Reasonable amount of outcomes. Proper CV and good no excluded patients.

**Overall Risk of bias**

Low