**PROBAST**

Study:

Combining patient-reported outcome measures to screen for active disease in rheumatoid arthritis and psoriatic arthritis.

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?**

Mixed effects regression model

**What is the outcome of interest?**

Active disease detection

Step 3: Assess risk of bias

**Domain 1: Participants**

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

For this study, we included patients who participated in the ‘treatment in the Rotterdam Early Arthritis Cohort’ trial (tREACH, ISRCTN26791028), the ‘Tapering strategies in Rheumatoid Arthritis’ trial (TARA, NTR2754) or the ‘Dutch southwest Early Psoriatic Arthritis cohoRt’ (DEPAR).The tREACH trial was a stratified, single- blinded randomised controlled trial with a follow- up period of 5 years. The trial was carried out in eight rheumatology centres in the southwestern part of the Netherlands. Patient recruitment took place between July 2007 and April 2011. Disease- modifying anti- rheumatic drug (DMARD)- naïve early undifferentiated arthritis (UA) and RA patients with an arthritis in ≥1 joint and symptom duration of <1 year were included. RA diagnosis was based on fulfilment of the 1987 or 2010 classification criteria.17 18 In the tREACH trial, multiple initial treatment strategies were compared. Patients received either (1) methotrexate, including DMARD combination therapies with or without glucocorticoid bridging therapy; (2) hydroxychloroquine or (3) non- steroid anti- inflammatory drugs/glucocorticoids as initial treatment. The tREACH trial had a T2T management approach that was aimed at reaching LDA, defined as a DAS with 44 joints ≤2.4. Treatment was intensified until LDA was achieved. Medi-cation was tapered if patients were in sustained remission, defined as DAS ≤1.6 at two consecutive visits. If a flare occurred, defined as a DAS >2.4, full treatment was restarted according to the stage of the protocol. The full treatment protocol is described elsewhere.19The TARA trial was a single- blinded randomised controlled trial with a follow- up period of 2 years. The trial was carried out in 12 rheumatology centres in the southwestern part of the Netherlands. The patient recruitment took place between September 2011 and July 2016. RA patients with well- controlled disease, defined as a DAS ≤2.4 and swollen joint count (SJC) ≤1 at two consecutive visits within 3 months, who used both ≥1 conventional synthetic (cs) DMARDs and a TNF- inhibitor, were included. RA diagnosis was based on fulfilment of the 1987 or 2010 classification criteria.1

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

N

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

N

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

High

**Rationale of bias rating**

There is considerable bias as patients were recruited from three different RCTs each with different eligibility criteria. Also patients were actively treated which makes the results not generalizable to most other RA patients.

**Domain 2: Predictors**

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

In all studies, patients completed questionnaires, including the PROMs selected from each ICHOM- recommended domain. These PROMs are general health/PGA (VAS, 0–100 mm), HAQ- DI, EQ- 5D, pain (VAS, 0–100 mm/Numeric Rating Scale (NRS), 0–10), fatigue (VAS, 0–100 mm/NRS, 0–10) and presenteeism (0%–100% productivity loss).

The change in PROM scores between two consecutive visits was categorised into two groups: ‘deterioration’ (1) and ‘no deterioration’ (0). For example, if the HAQ- DI score increased over a 3- month interval, indicating increased functional impair-ment, it was labelled as ‘deterioration’. If the HAQ- DI score remained the same or decreased, it was categorised as ‘no deterioration’. For the EQ- 5D and HAQ- DI, any worsening in score was considered a deterioration. For the PROMs that use a (transformed) 0–100 scale, a wors-ening of ≥10 points was defined as deterioration. This threshold was chosen to ensure consistent cut- off points between the VAS and transformed NRS.All dichotomised PROM changes were then included as potential explanatory variables in a regression model, with active disease as the dependent variable.

**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**

PROMs were assessed the same way for everyone.

**Domain 3: Outcome**

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

Disease activity is usually objectified by a physician in the outpatient clinic using a disease activity index. The composite disease activity measures were chosen because current T2T guidelines recommend the use of the DAS for RA and the DAPSA for PsA to assess disease activity. In both early UA/RA and established RA, disease activity was assessed by the DAS with a 44- SJC, tender joint count (TJC) measured with the Ritchie Articular Index in 53 joints, erythrocyte sedimentation rate (in millimetres (mm)/hour) and general health (VAS, 0–100 mm) or PGA (VAS, 0–100 mm).24 In addition, anti- citrullinated protein antibody (ACPA) and rheumatoid factor (RF) status were determined as well as the C reactive protein (CRP, in milligrams/decilitres) for the RA patients. In PsA, disease activity was measured with the DAPSA score, which is the sum of the 66- SJC, 68- TJC, CRP (in milligrams/decilitres), general health (VAS, 0–100 mm divided by 10) and pain (VAS, 0–100 mm divided by 10).

We chose the DAPSA as this is a continuous disease activity measure comparable to the DAS. In addition, the number of patients with arthritis, psoriasis and enthesitis was determined together with the body surface area (BSA, 0%–100%) and the Leeds Enthesitis Index (LEI, 0–6).

**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?**

PY

**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**

Active disease diagnosis was different between PSA and RA patients. However, because the cohorts were also analysed separately there is no bias risk here.

**Domain 4: Analysis**

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

From the 587 early UA/RA and 189 established RA patients, 495 and 188 had disease activity and PROM data at the same consecutive time points available, respec-tively. From the PsA cohort, 525 patients were included.

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

Mixed effects logistic regression models with an unstructured variance- covariance structure of the random effects and a robust sandwich covariate estimator were used. The significant variables were identified using backward elimination. Variables with a p>0.05 were removed from the model. To account for repeated measurements within one patient, a random intercept at the patient level was used. The selected PROMs from the regression analysis were combined to form a summarised score: PROM1+ PROM2+ PROM3+ ...PROMn . This score indicates the number of deteriorated PROMs. For example, a score of 0 means that none of the selected PROMs deteriorated, a score of 2 means that two PROMs deteriorated.

**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)**

Bootstrapping was used for internal validation of the diagnostic accuracy of the main model. All analyses were conducted separately for RA and PsA.

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

AUC, SEN, SPE

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

From the 587 early UA/RA and 189 established RA patients, 495 and 188 had disease activity and PROM data at the same consecutive time points available, respec-tively. From the PsA cohort, 525 patients were included.

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

If general health/PGA was missing, we used the three- item DAS to determine a patient’s disease status. This was the case in 57 out of 4594 RA visits. To determine if there was a difference in response between patients with a well- controlled disease and those with an active disease, we analysed missing PROM scores separately for each group. As shown in online supplemental table S1, the percentage of missing PROM scores varied between PROMs, ranging from 0% to 1% for PGA to 67% for productivity loss. In RA, missingness of PROM scores was similar between visits with a well- controlled and active disease, while in PsA missingness was 6%–7% higher in visits where patients developed a disease flare compared with those who continued having a well- controlled disease. Missing values were treated as missing and were not included in the analysis. All analyses were performed by using Stata V.18 (StataCorp).

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

Y

**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?**

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**

High

**Rationale of bias rating**

Unclear how many patients were excluded and what the criteria were for it. Only AUC score reported.

**Overall Risk of bias**

High