Development and validation of a model to predict the disease activity score: towards a remote treat-to-target approach for rheumatoid arthritis

Looijen, A.E.M.; Welsing, P.M.J.; Bergstra, S.A.; Helm-van Mil, A.H.M. van der; Jong, P.H.P. de

Ann Rheum Dis 85 (2026) 620–629

Contents lists available at ScienceDirect

Annals of the Rheumatic Diseases

journal homepage: https://www.sciencedirect.com/journal/annals-of-the-rheumatic-diseases

Rheumatoid arthritis

Development and validation of a model to predict the disease activity score: towards a remote treat-to-target approach for rheumatoid arthritis

Agnes E. M. Looijen[1,*], Paco M. J. Welsing[2], Sytske Anne Bergstra[3], Annette H. M. van der Helm-van Mil[1,3], Pascal H. P. de Jong[1]

[1] *Department of Rheumatology, Erasmus MC, Rotterdam, The Netherlands*
[2] *Department of Rheumatology & Clinical Immunology, University Medical Centre Utrecht, Utrecht, The Netherlands*
[3] *Department of Rheumatology, Leiden University Medical Center, Leiden, The Netherlands*

ARTICLE INFO

Article history:
Received 30 September 2025
Received in revised form 17 November 2025
Accepted 22 November 2025

ABSTRACT

Objectives: Remote monitoring of disease activity in patients with rheumatoid arthritis (RA) offers a promising solution to increasing healthcare demands. This study aimed to develop and validate a model using selected clinical and patient-reported outcome measure (PROM) items that efficiently and accurately reflect the original disease activity score (DAS).

Methods: Data from 5802 visits of 612 patients with RA from the treatment in the Rotterdam Early Arthritis Cohort and TApering strategies in RA trials were randomly split (1:1) into derivation and internal validation sets. An external validation was performed using 4404 visits from 1554 patients with RA from the Early Arthritis Cohort. A model was developed using Least Absolute Shrinkage and Selection Operator (LASSO) regression that incorporated age, sex, disease duration, autoantibody status, and individual PROM items, including visual analogue scale (VAS) general health, all Health Assessment Questionnaire-Disability Index (HAQ-DI) items, VAS pain, and VAS fatigue, to predict the DAS. The model's ability to detect active disease (DAS >2.4) and remission (DAS <1.6) was evaluated using the area under the receiver operating characteristic curve (AUC-ROC), along with sensitivity and specificity across predefined thresholds.

Results: The final model included 12 out of 28 predictors: age, sex, disease duration, VAS general health, 7 HAQ-DI items, and VAS pain. It showed excellent discriminative ability for detecting active disease with AUC-ROC values of 0.89 in both the development and internal validation sets, and 0.82 in the external validation set. For detecting remission, the AUC-ROC values were 0.86, 0.85, and 0.82, respectively. Test characteristics were provided for different thresholds.

Conclusions: The proposed DAS intended for digital remote assessment combines clinical and PROM items and can accurately and efficiently distinguish between disease activity states in RA, supporting its potential use in remote monitoring in the future.

*Correspondence to Agnes E. M. Looijen, Department of Rheumatology, Erasmus MC, Rotterdam, The Netherlands.
E-mail address: a.e.m.looijen@erasmusmc.nl (A.E.M. Looijen).
Handling editor Josef S. Smolen.

https://doi.org/10.1016/j.ard.2025.11.019

WHAT IS ALREADY KNOWN ON THIS TOPIC

- With rising demands on our healthcare system, remote monitoring can help sustain the quality of rheumatologic care by facilitating home-based detection of active disease and remission, which is essential for the recommended treat-to-target approach.
- The use of patient-reported outcome measures (PROMs) could help facilitate remote monitoring.
- However, current PROMs do not optimally align with the disease activity score, highlighting the need for improvement before they can be used to facilitate remote monitoring.

WHAT THIS STUDY ADDS

- A combination of age, sex, disease duration, visual analogue scale (VAS) general health, 7 Health Assessment Questionnaire-Disability Index items, and VAS pain can adequately differentiate between well-controlled and active disease with excellent diagnostic accuracy in an independent real-world cohort.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

- The digital disease activity score (digital DAS) combines 9 individual PROM items, thereby reducing questionnaire burden, which is sufficient for effectively detecting active disease and remission according to the DAS.
- These results may facilitate remote monitoring of patients with rheumatoid arthritis and aid in maintaining the quality of rheumatic care in the future.

INTRODUCTION

Current guidelines for rheumatoid arthritis (RA) recommend a treat-to-target (T2T) approach, which involves modifying treatment until remission or low disease activity (LDA) is achieved [1–3]. Moreover, if patients are in sustained remission, tapering of treatment can be considered [2,3]. For a T2T approach, it is essential to monitor disease activity using validated disease activity indices such as the disease activity score (DAS), which necessitates a physical examination. These assessments are recommended every 1-6 months depending on the disease activity level [1]. However, regular in-clinic assessments are increasingly challenging due to the growing demands on our healthcare system [4,5]. For example, previous literature showed that only 34% of patients with RA with active disease were assessed within the recommended 3-month timeframe [2,4,6]. Therefore, alternative methods to assess disease activity are necessary to uphold the quality of rheumatic care [5,7].

One option is remote monitoring using patient-reported outcome measures (PROMs) [4,5]. PROM usage is already recommended for daily practice as they help quantify disease impact [8–10]. When PROMs are used as a screening tool to identify disease activity, they may also help reduce the number of outpatient clinic visits [4,5]. Several studies have examined the use of PROMs to monitor disease activity, including PRO-CLinical ARthritis Activity (PRO-CLARA) and routine assessment of patient index data 3 (RAPID3), which are also recommended by the American College of Rheumatology for use in daily practice [11–14]. Although informative during clinic visits, their ability to accurately align with DAS scores remains moderate as the reported Spearman correlation coefficients between the DAS28-erythrocyte sedimentation rate (ESR) and the PRO-CLARA and RAPID3 were 0.57 and 0.66, respectively [15]. A combination of multiple dichotomised PROMs resulted in a moderate diagnostic accuracy (area under the receiver operating characteristic curve [AUC-ROC] of 0.76) for detecting active disease [16]. However, total PROM questionnaires may contain items unrelated to disease activity, limiting their discriminative ability. From the patient's perspective, answering a large number of questions can be burdensome and may negatively affect compliance [17]. Moreover, the aforementioned model lacks external validation, which is crucial for assessing its generalisability and clinical applicability.

Therefore, we aimed to identify and validate a subset of clinical and independent PROM items that most effectively reflect the DAS, enabling efficient and accurate discrimination between disease activity states in RA. This could be a step towards remote monitoring of RA.

METHODS

Patients and study data

For model development and internal validation, we included patients who participated in the 'treatment in the Rotterdam Early Arthritis Cohort' trial (tREACH, ISRCTN26791028) and the 'TApering strategies in Rheumatoid Arthritis' trial' (TARA, NTR2754). For external validation of the model, patients from the 'Early Arthritis Cohort' (EAC) were included. All study protocols were approved by ethics committees at each participating centre, and all patients provided written informed consent before inclusion in accordance with the Declaration of Helsinki.

The tREACH trial was a single-blinded randomised controlled trial (RCT) conducted across 8 rheumatology centres in the Netherlands. Patients were recruited between July 2007 and April 2011. Disease-modifying antirheumatic drug (DMARD)-naïve early undifferentiated arthritis and patients with RA, who had arthritis in ≥1 joint(s) and a symptom duration of <1 year, were included. RA diagnosis was based on fulfilment of the 1987 or 2010 classification criteria [18,19]. The trial compared multiple initial treatment strategies and had a T2T management approach that aimed for LDA, defined as a DAS ≤2.4. Treatment was intensified until LDA was achieved and tapered if patients were in sustained remission, defined as a DAS <1.6 at 2 consecutive visits. If a flare occurred, defined as a DAS >2.4, full treatment was restarted according to the stage of the protocol. Visits occurred every 3 months for 3 years and thereafter annually up to 5 years. More information on the tREACH trial can be found elsewhere [20,21].

The TARA trial was a single-blinded RCT conducted across 12 rheumatology centres in the Netherlands. Patients were included 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 2 consecutive visits, who used both ≥1 conventional synthetic (cs)DMARDs and a tumour necrosis factor (TNF) inhibitor, were included. RA diagnosis was based on fulfilment of the 1987 or 2010 classification criteria [18,19]. The trial compared 2 tapering strategies: tapering csDMARDs in year 1 followed by TNF-inhibitors in year 2, or vice versa. If a disease flare occurred, defined as DAS >2.4 or SJC >1, the last effective treatment was restarted and intensified until well-controlled disease was re-established. Patients were followed for 2 years, with scheduled visits occurring every 3 months. Extensive information on the TARA trial can be found elsewhere [22].

The EAC is an ongoing population-based inception cohort in Leiden, The Netherlands. In the EAC, consecutive patients with recent-onset arthritis (symptom duration <2 years) have been

included since 1993. RA diagnosis was based on fulfilment of the 1987 or 2010 classification criteria [18,19]. Treatment of included patients follows current (inter)national guidelines and is determined by the treating rheumatologist. Follow-up visits take place biannually in the first year and annually thereafter. Further details on the EAC are available elsewhere [23]. For this study, only patients with RA enrolled from January 2006 onwards were included as they were treated according to the currently recommended T2T principles [24]. The most recent dataset entry is from October 2023.

Included data for this analysis

In all cohorts, the DAS and its components were collected at each visit by a trained research nurse. PROMs were completed at home using paper or digital questionnaires before the scheduled appointment. For model development and internal validation, we included patients from the tREACH and TARA cohorts who met the 1987 or 2010 RA criteria and had ≥1 variable of interest and a concurrent DAS measurement. For external validation, all EAC visits meeting the RA criteria were included if all variables of the developed model, as well as the DAS, were available at the same visit. Due to the scope of the initial study cohorts, measurements of the outcome and predictors were not blinded from each other.

Outcome variable—disease activity

Disease activity was assessed with the DAS that includes a 44-SJC, a tender joint count measured with the Ritchie Articular Index in 53 joints, ESR, and general health (visual analogue scale [VAS], 0-100 mm) or patient global assessment (PGA) (VAS, 0-100 mm) [25].

Patient characteristics and clinical outcomes

At baseline, sex, age, symptom duration, disease duration, anticitrullinated protein antibody (ACPA), and rheumatoid factor (RF) status were documented. C-reactive protein and ESR were measured at each visit. Sex, age, disease duration, and autoantibody status (ACPA and RF) were included as explanatory variables in our analyses because evidence suggests that these variables may influence disease activity in patients with RA [26–29].

PROMs

For our analyses, we selected PROMs that (i) aligned with the International Consortium for Health Outcome Measurement (ICHOM)–recommended domains, (ii) were available in the tREACH, TARA, and EAC cohorts, and (iii) were applicable to all patients, excluding, eg, work-related questionnaires [10]. Based on the aforementioned criteria, the following PROMs were selected: general health/PGA, Health Assessment Questionnaire-Disability Index (HAQ-DI), pain, and fatigue. Table 1 shows the collected PROMs per ICHOM outcome domain and cohort [10].

General health/PGA (VAS, 0-100 mm) was included as a PROM to capture the disease activity component as well as the health impact domain. Higher scores indicate a poorer perceived general health. Activity limitation was measured with the HAQ-DI, with a total score ranging from 0 to 3, and higher scores indicating more functional impairment [30–32]. Pain was measured with a numeric rating scale (NRS, 0-10) in the tREACH and TARA trial, whereas in the EAC, a VAS (0-100 mm) was used. Higher scores indicate greater pain severity. Lastly, fatigue was measured using a VAS (0-100 mm) in the tREACH and EAC. In the TARA, an NRS (0-10) was used. Higher scores represent more severe fatigue [33]. For comparability, all NRS were transformed to a 0 to 100 scale.

Statistical analysis

To account for differences in RA stage (early vs established RA), baseline DAS, and treatment protocols between the tREACH (n = 425) and TARA (n = 187) cohorts, all visits from both cohorts were merged into 1 dataset and thereafter randomly split (1:1) into development and internal validation sets. This ensured a balanced representation of early and established RA visits with an equal range in disease activity levels across both sets. It also accounted for patients who participated in both the tREACH and, thereafter, the TARA trial. Visits from the included patients with EAC (n = 1554) served as an external validation set.

Model development

A 'Least Absolute Shrinkage and Selection Operator' (LASSO) regression was used to develop the DAS intended for digital remote assessment (digital DAS hereafter). LASSO selects the most relevant variables from a set of potential predictor variables and thereby decreases overfitting, resulting in a more generalisable model. Potential predictors to estimate the DAS included age, sex, disease duration, ACPA, RF, and all individual PROM items from the following questionnaires: VAS general health/PGA (1 item), HAQ-DI (20 items), VAS pain (1 item), and VAS fatigue (1 item). For this analysis, the use of aids or help from others was not considered for individual HAQ-DI items. All items were treated as linear continuous variables to maintain an evenly distributed weight between different levels within 1 item. Age and disease duration were defined at the time of visit. ACPA and RF status were assessed at baseline and were then carried forward to subsequent visits. If general health or PGA values were missing in the tREACH or TARA datasets, the 3-item DAS was used to impute the DAS. This occurred in 46 of 5802 RA visits. Missing predictors were not imputed, and the final model included only visits with complete data for all selected variables.

Several methods for selecting λ were evaluated for goodness of fit, including cross-validation (CV), adaptive lasso–which

Table 1
PROMs per ICHOM-recommended outcome domain and cohort

Outcome domains [10]	tREACH	TARA	EAC
Disease activity component and health impact	VAS general health, 0-100 mm	VAS PGA,0-100 mm	VAS general health, 0-100 mm
Activity limitations	HAQ-DI	HAQ-DI	HAQ-DI
Pain	NRS, 0-10	NRS, 0-10	VAS, 0-100 mm
Fatigue	VAS, 0-100 mm	NRS, 0-10	VAS, 0-100 mm

EAC, Early Arthritis Cohort; HAQ-DI, Health Assessment Questionnaire-Disability Index; ICHOM, International Consortium for Health Outcome Measurement; NRS, numeric rating scale; PGA, patient global assessment of disease activity; PROM, patient-reported outcome measure; TARA, TApering strategies in Rheumatoid Arthritis; tREACH, treatment in the Rotterdam Early Arthritis Cohort; VAS, visual analogue scale.

iteratively applies lasso with CV, removing variables with zero coefficients and reweighting the remaining ones in each step—and selection of λ based on the Bayesian Information Criterion (BIC) [34]. The BIC method was chosen for its optimal balance of model simplicity, lowest mean squared error, and highest R-squared. The selected items and their penalised coefficients were combined into the digital DAS as follows:

$$digital\ DAS = \beta_0 + \beta_1 \cdot item_1 + \beta_2 \cdot item_2 + \beta_3 \cdot item_3 + \ldots \beta_n \cdot item_n$$

Model fit evaluation and recalibration

Agreement between the digital DAS and DAS was evaluated using Bland-Altman plots [35,36]. Model calibration was evaluated by comparing the predicted probabilities for active disease (DAS >2.4) and remission (DAS <1.6) against the actual outcome (active disease or remission) [37,38]. Calibration plots were generated using logistic regression with an intercept of 0 and an offset coefficient for the digital DAS score as a predictive variable. Based on the calibration plot of active disease, the model was recalibrated using a regression-derived intercept and coefficient. The recalibrated digital DAS was re-evaluated in the development set using Bland-Altman and calibration plots.

Evaluation of discriminative ability

The digital DAS was evaluated for its ability to discriminate between well-controlled disease vs active disease and remission vs no remission using the AUC-ROC [38]. An AUC-ROC of 0.5 to 0.6 indicates no discriminative ability, 0.6 to 0.7 is considered poor, 0.7 to 0.8 is acceptable, 0.8 to 0.9 is excellent, and an AUC-ROC greater than 0.9 is considered outstanding [39,40]. Additionally, sensitivity and specificity were evaluated using thresholds based upon (i) the Youden index, which balances the sensitivity and specificity; and (ii-iii) 95% sensitivity and specificity in the development set. To identify misclassification patterns, patient characteristics of the external validation visits were shown for each cell of the confusion matrix (ie, correct and misclassified visits), stratified for the aforementioned thresholds.

Sensitivity analysis

A sensitivity analysis was conducted to compare the performance of the digital DAS with an alternative model that incorporated its complete PROM scores. This alternative model, referred to as the 'full digital DAS', included the total scores of each PROM selected in the digital DAS instead of the individual items. For instance, if multiple individual items from the HAQ-DI were selected, the overall HAQ-DI score was used in the full digital DAS as 1 single predictor. In the full digital DAS, the PROM scores—along with other selected clinical variables such as age and sex—were included as explanatory variables in a ridge regression model to estimate the DAS. The resulting penalised coefficients and variables were combined into a single formula. The evaluation of the model fit, recalibration, and assessment of the discriminative ability was done in the same manner as the main analysis.

RESULTS

Patient characteristics

The patient selection flowchart and baseline characteristics of the initial cohorts are shown in Supplementary Figure S1 and Supplementary Table S1, respectively. Patient characteristics of the visits used in the development, internal, and external validation sets are shown in Table 2. The development set included 2901 visits, and active disease was present in 564 (19%) of these visits. The median (IQR) disease duration was 2.0 (0.8-4) years, and the mean DAS (SD) was 1.6 (1.0). In 59% of the visits, ACPA was present, whereas RF was present in 57% of them.

The internal validation set included 2901 visits, and active disease was present in 576 (20%) of these visits. The median (IQR) disease duration was 1.8 (0.8-4.0) years. The mean DAS (SD) was 1.7 (1.0). In 58% and 59% of the visits, ACPA and/or RF were present, respectively.

The external validation set included 4404 visits. Active disease was present in 1599 (36%) of these visits. The median (IQR) disease duration was 2.0 (0.3-5.0) years, and mean (SD) DAS was 2.0 (1.3). ACPA and patients with RF positive were evaluated in 46% and 45% of the visits, respectively.

Model development, evaluation, and recalibration

The number of missing values in the development and internal validation set varied from no missingness to 27% of missing values (Supplementary Table S2). Ranges of the items were comparable between all 3 datasets (Supplementary Table S2). The selected model included 12 of the 28 items: age, sex, disease

Table 2
Characteristics at the visits of the development, internal, and external validation sets

Patient characteristics	Development (2901 visits)		Internal validation (2901 visits)		External validation (4404 visits)	
Sex, female, visits (%)	1914	(67)	1888	(65)	2735	(62)
Age (y), mean (SD)	56	(13)	56	(14)	60	(14)
ACPA positive,[a] visits (%)	1712	(59)	1695	(58)	1970	(46)
RF positive,[a] visits (%)	1667	(57)	1710	(59)	792	(45)
Disease duration (y)	2.0	(0.8-4.0)	1.8	(0.8-4.0)	2.0	(0.3-5.0)
DAS, mean (SD)	1.6	(1.0)	1.7	(1.0)	2.0	(1.3)
44 swollen joint count	0	(0-2)	0	(0-2)	0	(0-3)
53 tender joint count	1	(0-4)	1	(0-4)	2	(0-7)
ESR (mm/h)	11	(5-20)	11	(5-20)	11	(6-28)
CRP (mg/L)	4	(1-9)	4	(1-8)	3	(3-8)
VAS general health/PGA (0-100 mm)	23	(10-42)	23	(10-43)	30	(20-50)

ACPA, anticitrullinated protein antibody; CRP, C-reactive protein; DAS, disease activity score (measured in 44 joints); ESR, erythrocyte sedimentation rate; PGA, patient global assessment; RF, rheumatoid factor; VAS, visual analogue scale.
One patient can have multiple visits that are randomly split across the development and internal validation sets. This table shows the patient characteristics at the visits per set. Baseline characteristics of each initial cohort are shown in Supplementary Table S1. All results are presented as medians (IQR) unless indicated otherwise.
[a] Measured at baseline.

Table 3
Selected items and penalised coefficients for the digital DAS, before and after recalibration

Variable	Weight	Recalibrated weight
Age (y)[a]	0.002	0.006
Sex (male = 0, female = 1)	0.116	0.335
Diagnosis duration (y)[a]	−0.054	−0.155
VAS general health/patient global assessment (0-100 mm)	0.018	0.053
VAS pain (0-100 mm)	0.005	0.015
HAQ-DI rising: ability to get in and out of bed (0-3)	0.065	0.188
HAQ-DI eating: ability to cut your meat (0-3)	0.035	0.102
HAQ-DI eating: ability to open a new carton of milk (0-3)	0.017	0.049
HAQ-DI walking: ability to climb up 5 stairs (0-3)	0.007	0.021
HAQ-DI reach: ability to reach and get down a 5-lb object (0-3)	0.034	0.098
HAQ-DI grip: ability to open jars which have been previously opened (0-3)	0.056	0.160
HAQ-DI grip: ability to turn taps on and off (0-3)	0.017	0.048
Constant	0.814	−4.480

HAQ-DI, Health Assessment Questionnaire-Disability Index; digital DAS, disease activity score intended for digital remote assessment; VAS, visual analogue scale.
[a] Diagnosis duration and age were defined at the time of the visit.

duration, VAS general health/PGA, 7 HAQ-DI items, and VAS pain. Table 3 shows the selected items with their penalised coefficients. The selected HAQ-DI items covered the following categories: rising, eating, walking, reach, and grip.

After recalibration, the mean difference (95% limits of agreement) between the digital DAS and DAS was 3.8 (1.2, 6.4). In general, the DAS is higher compared to the digital DAS across all values (Supplementary Fig S2). The recalibrated model showed a good calibration for active disease, and moderate calibration for remission, with an overestimation of patients in remission (DAS <1.6, Supplementary Fig S3).

Discriminative ability

Across all datasets, the model demonstrated excellent discriminative ability in distinguishing between well-controlled vs active disease and remission vs no remission (Fig 1). In the development set, the AUC-ROC was 0.89 for well-controlled vs active disease and 0.86 for remission vs no remission. These findings were consistent in the internal validation set, with AUC-ROCs of 0.89 and 0.85, respectively. In the external validation set, the model maintained excellent discriminative ability, achieving an AUC-ROC of 0.82 for both comparisons.

Thresholds

Different thresholds can be chosen, depending on whether a sensitive or specific measure would be preferred, or a balance herein. To investigate the effect of different thresholds on sensitivity and specificity, 3 thresholds were determined based on data from the development set: (i) the Youden index, which equally balances sensitivity and specificity; and (ii-iii) thresholds achieving 95% sensitivity and specificity, respectively. As shown in Figure 2, the sensitivity and specificity of all 3 thresholds are highest in the development set and show a slight decline when the same thresholds are applied in the internal and external

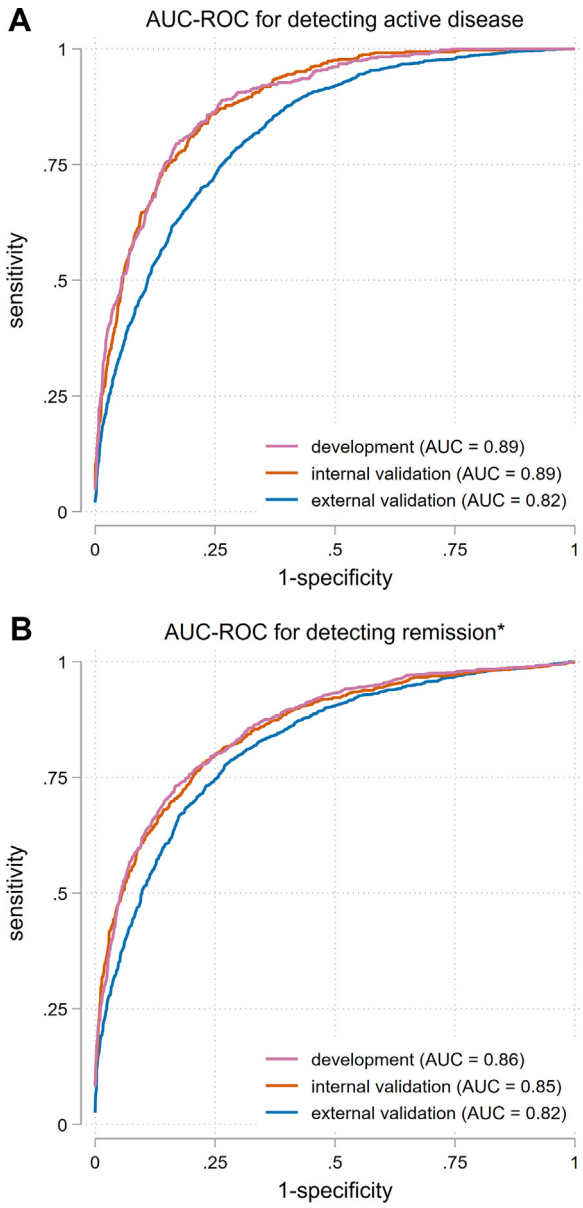

Figure 1. Receiver operating characteristic curve for detecting active disease (A) and remission (B) in the development, internal, and external validation set. *The graph shows the ability of the digital DAS to differentiate between a DAS <1.6 and ≥1.6. AUC, area under the curve; AUC-ROC, area under the receiver operating characteristic curve; DAS, disease activity score; digital DAS, disease activity score intended for digital remote assessment.

validation sets. Using the Youden index (threshold ≥ −1.27), the external validation yields a sensitivity of 66% and specificity of 81% for detecting active disease (Fig 2A). When applying the threshold that achieves 95% sensitivity (≥ −2.77) in the development set, then the sensitivity and specificity in the external validation set are 91% and 51%, respectively. Conversely, the threshold optimised for 95% specificity (≥0.15) yields 34% sensitivity and 95% specificity in the external validation set. The same interpretation can be applied for remission (Fig 2B).

Remote monitoring is useful when it leads to a substantial reduction in patient visits, whereas patients with active disease are not overlooked. In practice, the number of patients with active and well-controlled disease can vary, which can impact the benefit of the digital DAS. To illustrate this, we evaluated the correct and incorrect classifications in the real-world cohort (external validation cohort). As shown in Figure 3A, in 36% of

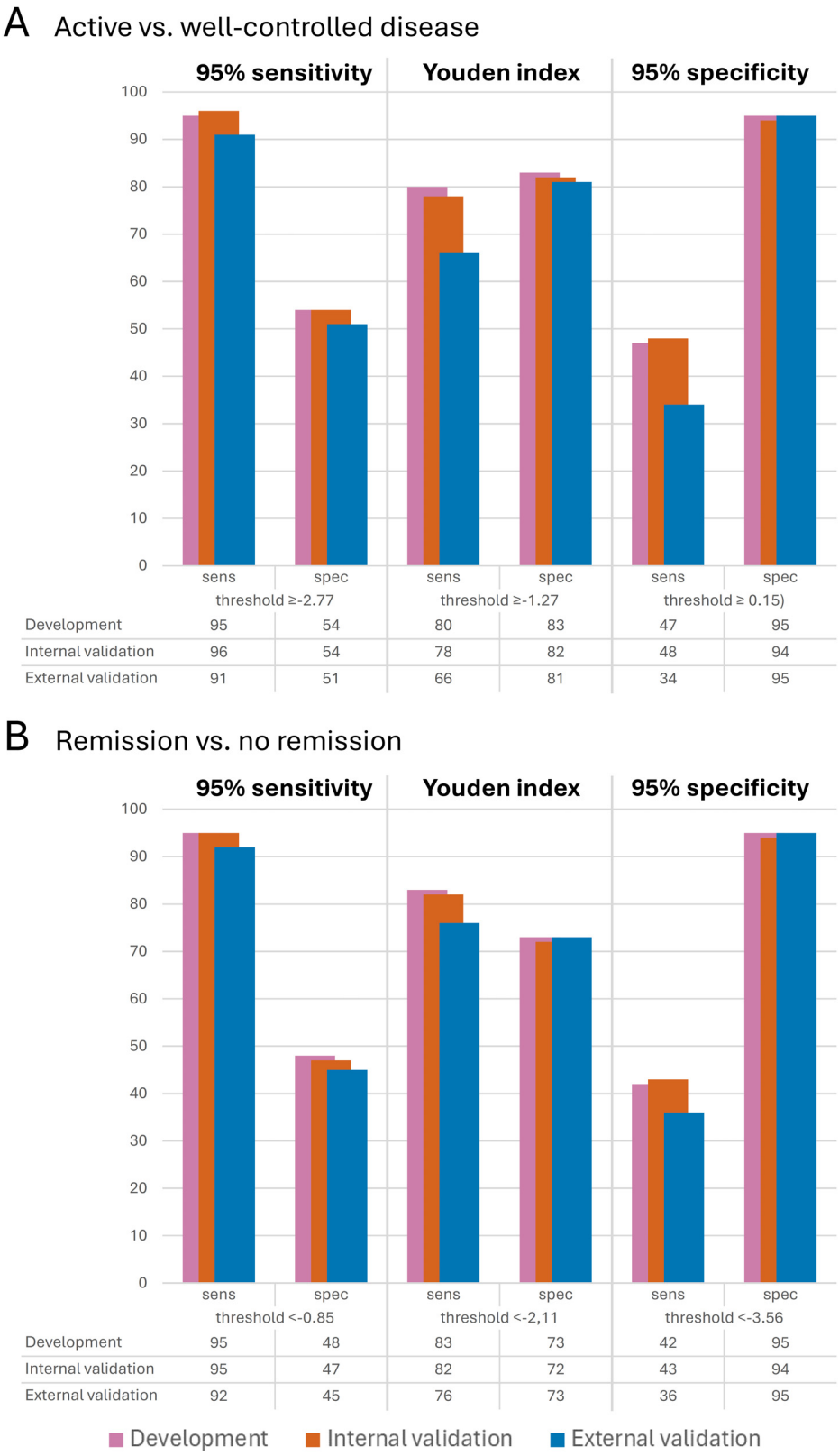

Figure 2. Sensitivity and specificity for different thresholds of the digital DAS in the development, internal, and external validation set for active disease (A) and remission (B). The thresholds are determined in the development set and are thereafter applied to all 3 datasets. Digital DAS, disease activity score intended for digital remote assessment.

the visits, patients had active disease. Using the Youden index (≥ −1.27), 24% of visits were correctly classified as active disease, 12% actually had active disease and were misclassified as well-controlled disease, 12% actually had well-controlled disease and were misclassified as active disease, and 51% were correctly classified as well-controlled disease. The same was done applying a 95%-sensitivity threshold (≥ −2.77). In this case, 33% of the visits were correctly identified as active disease, only 3% of visits with active disease were misclassified as well-controlled disease, 31% of the visits were false positives, and 33% of the visits were correctly classified as well-controlled disease. Conversely, using a 95%-specificity threshold (≥0.15) resulted in 60% of visits being correctly classified as well-controlled disease, but 24% of active disease visits were missed. For remission,

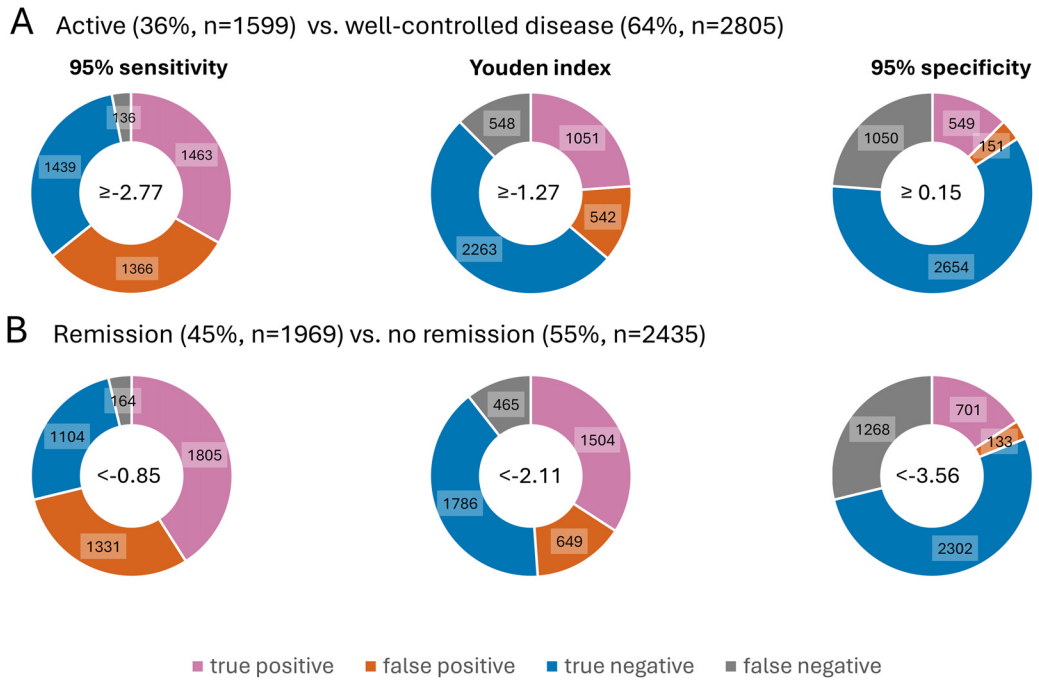

Figure 3. Distribution of visits in the external validation set across each cell of the confusion matrix based on the predefined thresholds for active disease (A) and remission (B). The figure shows the number of visits in the external validation set, classified by the model as follows: (i) correctly identified as active disease (A) or remission (B)—true positives (pink); (ii) incorrectly identified as active disease (A) or remission (B)—false positives (orange); (iii) correctly identified as well-controlled disease (A) or no remission (B)—true negatives (blue); or (iv) incorrectly identified as well-controlled disease (A) or no remission (B)—false negatives (grey).

the same interpretation applies, and its classification distribution is illustrated in Figure 3B.

Thus, inherent to the chosen threshold, choosing a sensitive definition results in very few visits with active disease that were missed, but at the cost of more false positives (patients with well-controlled disease that were identified as having active disease). The opposite was the case when choosing a very specific threshold. When using the Youden index as threshold, the percentage of visits with false-positive and false-negative DAS classifications was balanced (both 12%).

Characteristics of the visits per (mis)classification

To better understand the consequences of misclassifications, we studied the patient characteristics of the correct and misclassified visits. Supplementary Table S3 provides the patient characteristics for true-positive, true-negative, false-positive, and false-negative visits. In general, misclassified patients show a discordance of their VAS general health/PGA with their SJC and ESR. For example, in the case of active disease, patients with a false-negative test result had a lower VAS general health/PGA compared to the true positives, whereas no strong differences in SJC or ESR were observed between true positives and false negatives. Conversely, patients with a false-positive test result had a higher VAS general health/PGA compared to the true negatives, with no big observed differences in SJC or ESR between both groups. Thus, digital DAS misclassifications were often explained by discordance within the DAS between patient-reported components such as VAS general health/PGA and the more inflammation-based components, such as SJC and ESR.

Sensitivity analysis

A sensitivity analysis was conducted to evaluate the difference between a model composed of full PROM scores using a total of 27 individual items compared to the digital DAS using 12 individual items. Supplementary Table S4 and Supplementary Figs S4 and S5 show the coefficients, model fit and calibration plots of the full digital DAS. The ability to discriminate between well-controlled disease and active disease was similar for the full digital DAS compared to the digital DAS (AUC-ROC 0.84 vs 0.82 in the external validation set, Supplementary Fig S6A). Similar results were also found for remission vs no remission (AUC-ROC 0.84 vs 0.82 in the external validation set, Supplementary Fig S6B). Supplementary Figure S7 shows the sensitivity and specificity across different thresholds.

DISCUSSION

As demands on our healthcare system continue to rise, the implementation of remote monitoring using PROMs could help sustain the quality of our rheumatic care, including its T2T approach. Therefore, we assessed whether a model based on several clinical characteristics and selected PROM items could accurately reflect the DAS and its disease activity states in RA. Using LASSO regression, 12 of 28 possible explanatory items were combined. This resulted in a proposed DAS intended for digital remote assessment (digital DAS), which showed excellent discriminative ability with an externally validated AUC-ROC of 0.82 for both active disease and remission. Sensitivity and specificity varied depending on predefined thresholds. These results could be a step towards efficient remote monitoring with PROMs.

We have not chosen 1 specific threshold, allowing potential users to select thresholds based on their preference for high sensitivity, high specificity, or a balanced trade-off between the 2. To demonstrate the impact of these choices, we evaluated true and false classifications within a real-world RA cohort, where 36% of the visits entailed active disease and 64% well-controlled disease. A high-sensitivity threshold correctly identified 91% of

these active disease visits. Additionally, 51% of visits with well-controlled disease were correctly classified as such. If remote monitoring was used to reduce the need for patients with well-controlled disease to come to the rheumatology clinic, these patients could skip their physical outpatient visit. In our real-world cohort, this would result in a reduction of 36% of visits. A threshold with a balance between sensitivity and specificity correctly specified 66% of visits with active disease and enabled a 63% reduction in total visits. A high-specificity threshold correctly identified only 34% of active disease visits, but allowed for an 84% reduction in total visits.

The choice of threshold not only affects the number of correctly classified active or well-controlled patients but also results in a different percentage of patients with active disease being missed or patients with well-controlled disease being classified as active. For example, although a high-sensitivity threshold misclassifies only 9% of active disease visits, a balanced or high-specificity threshold misses 34% and 66%, respectively. Patients who were misclassified by the digital DAS often had a discordance between patient-reported components of the DAS (eg, general health/PGA) and more inflammatory-based components (eg, SJC and ESR). This indicates a misalignment in patients' perceived disease burden and inflammatory activity. To reduce this misalignment, the model could be extended with other home measurements that are a proxy for joint inflammation, such as hand function or finger fold measurements [41,42]. It is also worth noting that the identification of patients with a high disease burden, even without objective inflammatory activity, may not be undesirable, as these individuals still require additional care. This is consistent with the proposed dual T2T strategy in which both disease activity and disease impact are addressed [9].

In addition to diagnostic accuracy, the burden on patients to complete questionnaires is an important consideration when implementing models for remote monitoring. Therefore, we aimed to develop a parsimonious model using as few PROM items as possible. The proposed digital DAS, which integrates 9 PROM items with 3 clinical variables (not patient reported), demonstrated comparable performance to a model using the full PROM scores, which used 24 PROM items. This substantial reduction in items contributes to a more user-friendly and feasible tool for clinical applicability.

Alongside diagnostic accuracy and lowering questionnaire burden, several other factors should be considered when implementing a PROM-based measure for remote monitoring, including health literacy and compliance to complete the questionnaires. [17]. Additionally, some patients may still prefer the traditional outpatient clinic visits. Nevertheless, remote monitoring has the potential and flexibility to accommodate and facilitate patients' health care needs, as it may lead to a significant reduction in the frequency of outpatient clinical visits for the entire RA population.

This is not the first publication on remote monitoring using PROMs. However, the here developed digital DAS outperformed our earlier model in distinguishing well-controlled from active disease in RA (AUC-ROC = 0.76) [16]. Compared to the RAPID3, the digital DAS had a similar AUC-ROC and sensitivity (AUC-ROC = 0.80 and sensitivity = 93% vs AUC-ROC = 0.82 and sensitivity = 91%, respectively), but the digital DAS has a higher specificity (39% vs 51%) [43]. As the prevalence of active disease is generally lower due to improved treatment strategies, the increase in specificity will significantly reduce the number of outpatient clinic visits, whereas most patients with active disease are detected [44].

Other questionnaires, such as the Rheumatoid Arthritis Flare Questionnaire, also have a good discriminative ability for detecting disease flares [45]. However, these questionnaires are based on a different flare definition, which integrates aspects from both the physician's and patient's point of view [45,46]. In contrast, the digital DAS aligns with DAS-defined disease states commonly used in daily practice. Thus, the digital DAS serves a different purpose in comparison to the aforementioned tools.

The strengths of our study include the inclusion of both patients with early and established RA from 2 RCTs, with a protocolised T2T management approach, for model development and internal validation. Additionally, a real-world, independent cohort was used for external validation, further enhancing the robustness and applicability of our findings. All 3 datasets also included a large sample size of patients and visits.

There are also several issues that warrant consideration. Although the digital DAS showed strong discriminative ability across disease activity states, it initially overestimated scores at low DAS levels and underestimated them at higher DAS levels Supplementary Fig S1A. Overestimation could be caused by a higher perceived disease burden without objective inflammation or lasting functional impairment due to joint damage [47]. Underestimation at higher DAS levels may be caused by a disbalance in active vs nonactive disease visits in the development set. Recalibration improved the predicted probability of active disease, but led to an overestimation of predicted probability at lower DAS levels, limiting optimal prediction of remission probability. This trade-off was considered acceptable given the clinical priority of detecting active disease.

It should also be noted that PROMs may be sensitive to the time period in which they are completed. Additionally, PROMs, such as the HAQ-DI, were not designed for single item-level extraction or for application in contemporary remote monitoring settings, but rather primarily for scientific research purposes. This may limit the model's applicability in current patient populations. To mitigate this limitation, we prioritised the inclusion of patients who were treated according to the currently recommended T2T approach and had early as well as established disease to ensure that the model remains as representative as possible, despite the temporal gap of the study population.

Lastly, minor cohort differences, such as different HAQ-DI versions and use of VAS general health or VAS PGA, may have affected model performance. Nonetheless, the digital DAS maintained excellent discriminative ability during external validation, affirming its robustness and potential applicability in broader settings.

Future research should focus on investigating other measurement properties, such as responsiveness over time within 1 individual [48]. The model should also be tested in a fully remote setting using 1 universal set of questions to confirm that its discriminative ability remains intact in this environment. Additionally, updating the model with more modern questions, along with incorporating variables that specifically reflect joint inflammation, could further improve the model.

In conclusion, the digital DAS, incorporating age, sex, disease duration, VAS general health/PGA, 7 HAQ-DI items, and VAS pain, showed excellent diagnostic accuracy in identifying active disease and remission, even after external validation in an independent, real-world cohort. Although further validation and optimisation are preferred in future research, these findings support its potential for efficient remote monitoring of patients with RA.

Statements

The manuscript was prepared following the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) statement [49,50] (see also Supplementary Table S5 for the completed checklist).

Acknowledgements

We gratefully thank all participating patients, rheumatologists, and research nurses.

Contributors

AEML performed the statistical analysis and drafted the first version of the manuscript. PMJW and PHPdJ contributed to the analysis. SAB supplied the EAC data. All authors contributed to the design, revised the manuscript, and read and approved the final manuscript. PHPdJ is the guarantor.

Funding

This study was supported by an unrestricted grant from ZonMW (grant number 1711 + 02014), Pfizer (grant number WI229707) and AbbVie B.V. (grant number SA-003854).

Competing interests

SAB received an ASPIRE grant from Pfizer and speaker fees from Benecke and reports a relationship with Leiden University Medical Center that includes funding grants and speaking and lecture fees. The other authors have declared no conflicts of interest.

Patient consent for publication

Not applicable.

Ethics approval

Written informed consent was obtained from all participants according to the Declaration of Helsinki. The tREACH and TARA studies were approved by the ethics committee of the Erasmus MC (MEC-2006-252 for tREACH and MEC-2011-141 for TARA). The EAC was approved by the Local Medical Ethics Committee, named 'Commissie Medische Ethiek' (B19.008).

Provenance and peer review

Not commissioned; externally peer reviewed.

Patient and public involvement

Patients and/or the public were involved in the design and conduct of this research.

Data availability statement

The data underlying this article will be shared on reasonable request to the corresponding author.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the first author used Microsoft Copilot in order to assist with spelling, punctuation, and grammar. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.ard.2025.11.019.

Orcid

Agnes E. M. Looijen: http://orcid.org/0000-0002-5280-2182
Paco M. J. Welsing: http://orcid.org/0000-0003-2361-2803
Sytske Anne Bergstra: http://orcid.org/0000-0002-7136-5248
Annette H. M. van der Helm-van Mil: http://orcid.org/0000-0001-8572-1437
Pascal H. P. de Jong: http://orcid.org/0000-0001-6628-6222

REFERENCES

[1] Smolen JS, Breedveld FC, Burmester GR, Bykerk V, Dougados M, Emery P, et al. Treating rheumatoid arthritis to target: 2014 update of the recommendations of an international task force. Ann Rheum Dis 2016;75(1):3–15.
[2] Smolen JS, Landewé RBM, Bergstra SA, Kerschbaumer A, Sepriano A, Aletaha D, et al. EULAR recommendations for the management of rheumatoid arthritis with synthetic and biological disease-modifying antirheumatic drugs: 2022 update. Ann Rheum Dis 2023;82(1):3–18.
[3] Fraenkel L, Bathon JM, England BR, St Clair EW, Arayssi T, Carandang K, et al. 2021 American College of Rheumatology guideline for the treatment of rheumatoid arthritis. Arthritis Care Res 2021;73(7):924–39.
[4] Dejaco C, Landewé RBM. Controversies in rheumatology: telemedicine-friend or foe? Rheumatology (Oxford) 2023;62:2661–4.
[5] Knitza J, Gupta L, Hügle T. Rheumatology in the digital health era: status quo and quo vadis? Nat Rev Rheumatol 2024;20(12):747–59.
[6] Barber CEH, Mosher D, Dowling S, Bohm V, Solbak NM, Macmullan P, et al. Implementation and evaluation of audit and feedback for monitoring treat-to-target (T2T) strategies in rheumatoid arthritis using performance measures. Rheumatol Ther 2020;7(4):909–25.
[7] De Thurah A, Bosch P, Marques A, Meissner Y, Mukhtyar CB, Knitza J, et al. 2022 EULAR points to consider for remote care in rheumatic and musculoskeletal diseases. Ann Rheum Dis 2022;81(8):1065–71.
[8] Black N. Patient reported outcome measures could help transform healthcare. BMJ 2013;346:f167.
[9] Ferreira RJO, Ndosi M, de Wit M, Santos EJF, Duarte C, Jacobs JWG, et al. Dual target strategy: a proposal to mitigate the risk of overtreatment and enhance patient satisfaction in rheumatoid arthritis. Ann Rheum Dis 2019;78(10):e109.
[10] Oude Voshaar MAH, Das Gupta Z, Bijlsma JWJ, Boonen A, Chau J, Courvoisier DS, et al. International consortium for health outcome measurement set of outcomes that matter to people living with inflammatory arthritis: consensus from an international working group. Arthritis Care Res 2019;71(12):1556–65.
[11] Pickles T, Macefield R, Aiyegbusi OL, Beecher C, Horton M, Christensen KB, et al. Patient reported outcome measures for rheumatoid arthritis disease activity: a systematic review following COSMIN guidelines. RMD Open 2022;8(1):e002093.
[12] Hendrikx J, de Jonge MJ, Fransen J, Kievit W, van Riel PL. Systematic review of patient-reported outcome measures (PROMs) for assessing disease activity in rheumatoid arthritis. RMD Open 2016;2(2):e000202.
[13] England BR, Tiong BK, Bergman MJ, Curtis JR, Kazi S, Mikuls TR, et al. 2019 Update of the American College of Rheumatology recommended rheumatoid arthritis disease activity measures. Arthritis Care Res 2019;71(12):1540–55.
[14] Salaffi F, Migliore A, Scarpellini M, Corsaro SM, Laganà B, Mozzani F, et al. Psychometric properties of an index of three patient reported outcome (PRO)

measures, termed the CLinical ARthritis Activity (PRO-CLARA) in patients with rheumatoid arthritis. The NEW INDICES study. Clin Exp Rheumatol 2010;28(2):186–200.
[15] Salaffi F, Di Carlo M, Farah S, Marotto D, Atzeni F, Sarzi-Puttini P. Rheumatoid arthritis disease activity assessment in routine care: performance of the most widely used composite disease activity indices and patient-reported outcome measures. Acta Biomed 2021;92(4):e2021238.
[16] Looijen AEM, Snoeck Henkemans SVJ, Van Der Helm-Van Mil AHM, Welsing PMJ, Koc GH, Luime JJ, et al. Combining patient-reported outcome measures to screen for active disease in rheumatoid arthritis and psoriatic arthritis. RMD Open 2024;10(4):e004687.
[17] Glenwright BG, Simmich J, Cottrell M, O'Leary SP, Sullivan C, Pole JD, Russell T. Facilitators and barriers to implementing electronic patient-reported outcome and experience measures in a health care setting: a systematic review. J Patient Rep Outcomes 2023;7(1):13.
[18] Arnett FC, Edworthy SM, Bloch DA, McShane DJ, Fries JF, Cooper NS, et al. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. Arthritis Rheum 1988;31(3):315–24.
[19] Aletaha D, Neogi T, Silman AJ, Funovits J, Felson DT, Bingham 3rd CO, et al. 2010 Rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. Arthritis Rheum 2010;62(9):2569–81.
[20] Claessen SJ, Hazes JM, Huisman MA, van Zeben D, Luime JJ, Weel AE. Use of risk stratification to target therapies in patients with recent onset arthritis; design of a prospective randomized multicenter controlled trial. BMC Musculoskelet Disord 2009;10:71.
[21] de Jong PH, Hazes JM, Han HK, Huisman M, van Zeben D, van der Lubbe PA, et al. Randomised comparison of initial triple DMARD therapy with methotrexate monotherapy in combination with low-dose glucocorticoid bridging therapy; 1-year data of the tREACH trial. Ann Rheum Dis 2014;73(7):1331–9.
[22] van Mulligen E, de Jong PHP, Kuijper TM, van der Ven M, Appels C, Bijkerk C, et al. Gradual tapering TNF inhibitors versus conventional synthetic DMARDs after achieving controlled disease in patients with rheumatoid arthritis: first-year results of the randomised controlled TARA study. Ann Rheum Dis 2019;78(6):746–53.
[23] de Rooy DP, van der Linden MP, Knevel R, Huizinga TW, van der Helm-van Mil AH. Predicting arthritis outcomes—what can be learned from the Leiden Early Arthritis Clinic? Rheumatology (Oxford) 2011;50(1):93–100.
[24] Matthijssen XME, Niemantsverdriet E, Huizinga TWJ, Van Der, Helm-Van Mil AHM. Enhanced treatment strategies and distinct disease outcomes among autoantibody-positive and -negative rheumatoid arthritis patients over 25 years: a longitudinal cohort study in the Netherlands. PLOS Med 2020;17(9):e1003296.
[25] Fransen J, van Riel PL. The Disease Activity Score and the EULAR response criteria. Rheum Dis Clin North Am 2009;35(4):745–57 vii-viii.
[26] Koc GH, Looijen AEM, Van Der Horst-Bruinsma IE, De Jong PHP. Sex differences in rheumatoid arthritis: new insights from clinical and patient-reported outcome perspectives. J Rheumatol 2025;52(6):553–62.
[27] Movahedi M, Cesta A, Li X, Bombardier C, investigators OBRI. Disease activity trajectories for early and established rheumatoid arthritis: real-world data from a rheumatoid arthritis cohort. PLoS One 2022;17(9):e0274264.
[28] Katchamart W, Koolvisoot A, Aromdee E, Chiowchanwesawakit P, Muengchan C. Associations of rheumatoid factor and anti-citrullinated peptide antibody with disease progression and treatment outcomes in patients with rheumatoid arthritis. Rheumatol Int 2015;35(10):1693–9.
[29] Innala L, Berglin E, Möller B, Ljung L, Smedby T, Södergren A, et al. Age at onset determines severity and choice of treatment in early rheumatoid arthritis: a prospective study. Arthritis Res Ther 2014;16(2):R94.
[30] Bruce B, Fries JF. The Health Assessment Questionnaire (HAQ). Clin Exp Rheumatol 2005;23(5 Suppl 39):S14–8.
[31] Boers M, Jacobs JW, van Vliet, Vlieland TP, van Riel PL. Consensus Dutch health assessment questionnaire. Ann Rheum Dis 2007;66(1):132–3.
[32] Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23(2):137–45.
[33] Strand V, Boers M, Idzerda L, Kirwan JR, Kvien TK, Tugwell PS, et al. It's good to feel better but it's better to feel good and even better to feel good as soon as possible for as long as possible. Response criteria and the importance of change at OMERACT 10. J Rheumatol 2011;38(8):1720–7.
[34] StataCorp. STATA LASSO reference manual. Texas: Stata Press; 2021.
[35] Giavarina D. Understanding Bland Altman analysis. Biochem Med 2015;25(2):141–51.
[36] Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res 1999;8(2):135–60.
[37] Dougados M, Aletaha D, van Riel P. Disease activity measures for rheumatoid arthritis. Clin Exp Rheumatol 2007;25(5 Suppl 46):S22–9.
[38] Dankers F.J.W.M., Traverso A., Wee L., et al. Prediction modeling methodology. 2018. In: Kubben P., Dumontier M., Dekker A., editors. Fundamentals of Clinical Data Science [Internet]. Cham (CH): Springer; 2019. Chapter 8. Available from: https://www.ncbi.nlm.nih.gov/books/NBK543534/
[39] Hosmer Jr. DW, Lemeshow S, Sturdivant RX. Applied logistic regression. Hoboken: John Wiley & Sons; 2013.
[40] Hoo ZH, Candlish J, Teare D. What is an ROC curve? Emerg Med J 2017;34(6):357–9.
[41] Krijbolder DI, Khidir SJH, Matthijssen XME, Ten Brinck RM, van Aken J, Speyer I, et al. Hand function is already reduced before RA development and reflects subclinical tenosynovitis. RMD Open 2023;9(1):e002885.
[42] Venerito V, Manigold T, Capodiferro M, Markham D, Blanchard M, Iannone F, et al. Single-camera motion capture of finger joint mobility as a digital biomarker for disease activity in rheumatoid arthritis. Rheumatol Adv Pract 2025;9(2):rkae143.
[43] Wiegel J, Seppen BF, Ter Wee MM, Nurmohamed MT, Boers M, Bos WH. The RAPID3 questionnaire as a screening tool to reduce the number of outpatient clinic visits: a retrospective cohort study. Clin Rheumatol 2022;41(8):2525–31.
[44] Scott IC, Ibrahim F, Panayi G, Cope AP, Garrood T, Vincent A, et al. The frequency of remission and low disease activity in patients with rheumatoid arthritis, and their ability to identify people with low disability and normal quality of life. Semin Arthritis Rheum 2019;49(1):20–6.
[45] Bingham CO, Alten R, Bartlett SJ, Bykerk VP, Brooks PM, Choy E, et al. Identifying preliminary domains to detect and measure rheumatoid arthritis flares: report of the OMERACT 10 RA Flare Workshop. J Rheumatol 2011;38(8):1751–8.
[46] Alten R, Pohl C, Choy EH, Christensen R, Furst DE, Hewlett SE, et al. Developing a construct to evaluate flares in rheumatoid arthritis: a conceptual report of the OMERACT RA Flare Definition Working Group. J Rheumatol 2011;38(8):1745–50.
[47] Bombardier C, Barbieri M, Parthan A, Zack DJ, Walker V, Macarios D, et al. The relationship between joint damage and functional disability in rheumatoid arthritis: a systematic review. Ann Rheum Dis 2012;71(6):836–44.
[48] Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. BMC Med Res Methodol 2010;10(1):22.
[49] Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med 2015;162(1):55–63.
[50] Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015;162(1):W1–73.